History of main

June 6, 2011

Supervised Topic Modeling in Clinical Text

Perotte A Li Y Pivovarov R Weiskopf N Wood F
Columbia University, New York, NY 10027, USA
{{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [4], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: International Challenge: Classifying Clinical Free Text Using Natural Language Processing (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Methods

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patients chief complaint, diagnostic findings, therapy administered, patients response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within

a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patients visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 Supervised Topic Models

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

2.4 Generative Model

Given a fixed set of parameters: $\beta_{1:K}$, the K topics, α, the parameter for the document specific dirichlet distribution, and η, the response variable parameter, the generative process for documents and responses is as follows:

1. Draw topic proportions $\theta \mid \alpha \sim Dir(\alpha)$
2. For each word:
   (a) Draw topic assignment $z_n \mid \theta \sim Mult(\theta)$
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$
3. For each ICD-9 code from the root of the hierarchy and recursively descending the tree:
   (a) Draw regression coefficient $\eta_i \mid \mu \sim \mathcal{N}(\mu, 1)$
   (b) Draw a response variable $y_i \mid \bar{z}, \eta_i, Y, \xi \sim \frac{\Phi(\eta_i^T \bar{z})\psi(Y)}{Z}$ where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\psi(Y) = \exp\left\{\sum_{\{p,c\}} f(p,c)\right\}$ and

$$\begin{cases} f(0,0) = \xi_0 \\ f(0,1) = -\infty \\ f(1,0) = \xi_1 \\ f(1,1) = \xi_2 \end{cases}$$

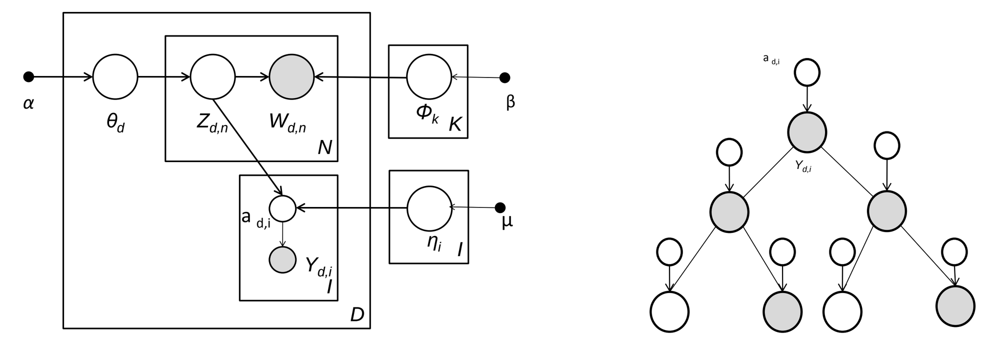


Figure 1: adapted sLDA model

We will employ a data augmentation scheme with auxiliary variables a_i in the probit model such that:

$$y_i \sim \begin{cases} 1, & a_i \geq 0 \\ 0, & a_i < 0 \end{cases} \tag{1}$$

$$a_i \sim \mathcal{N}\left(\eta_i^T \bar{z}, 1\right) \tag{2}$$

$$\eta_i \sim \mathcal{N}\left(\mu = 0, 1\right) \tag{3}$$

2.5 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation 4.

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: markov chain monte carlo (MCMC). Since in this model it is possible to sample from the conditional distributions for all variables we will use the gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a Rao-Blackwellized gibbs sampler for the supervised topic model for general exponential family response variables and then for our model with a probit response variable.

2.6 Rao-Blackwellization

To derive the Rao-Blackwellized gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right)}{\int_{\theta} p(\theta \mid \alpha)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right) d\theta} \tag{4}$$

$$\int_{\theta}\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta\, d\phi_{1:K} = \int_{\theta}\int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M}\left\{p(\theta_m; \alpha)\prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m)\, p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^{I} p(\eta_i; \mu)\, d\theta\, d\phi_{1:K} \tag{5}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_{\theta}\prod_{m=1}^{M} p(\theta_m; \alpha)\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)\, d\theta \tag{6}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{7}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^{M} h(y)\exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{8}$$

2.7 Gibbs Sampling

To derive the gibbs sampler for the general exponential family form as well as the probit regression model, we derive the individual conditional probability distributions for all latent variables and parameters.

2.7.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, m. The conditional probability with respect to this latent variable is proportional to the joint distribution up to a constant.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \propto p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \tag{9}$$

Due to the factorization of this model we can rewrite the joint distribution as the following:

$$\propto \prod_{i=1}^{I}\left[p(\eta_i; \mu)\, p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \tag{10}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i=1}^{I}\left[p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{11}$$

Here, $n_{m,v}^{k,-(m,n)}$ represents the count of word v in document m assigned to topic k omitting the $(m,n)^{th}$ word count. For exponential family distributions, the normalization contant, $h(y_{m,i})$, does not depend on $z_{m,n}$.

$$\propto \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{12}$$

Given this expression, $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right)$ can be sampled through enumeration as seen in Equation 13. In the case of probit regression, the expression for 12 evaluates to Equation 14. Equivalently we can parameterize the model with an auxiliary variable a_i, resulting in Equation 15.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) = \frac{\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}}{\sum_{k=1}^{K}\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}} \tag{13}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\Phi\left(\eta_i^T \bar{z}\right)^{y_{m,i}}\left(1 - \Phi\left(\eta_i^T \bar{z}\right)\right)^{1 - y_{m,i}}\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{14}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \mathbf{a}, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\delta\left(sign\left(a_{m,i}\right) = 2y_{m,i} - 1\right)\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{15}$$

2.7.2 $p\left(\eta_i \mid \mathbf{z}, \mathbf{Y}, \mu\right)$ or $p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right)$in the augmented probit regression model

Given that η_i and $a_{m,i}$ are distributed normally, this posterior distribution is also normal. In the case for general exponential family distributions, η_i can remain a parameter without a prior, fit with maximimum likelihood in the usual fashion.

$$p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right) = \mathcal{N}\left(\eta_i \mid \hat{\mu_i}, \hat{\mathbf{S}}_i\right) \tag{16}$$

$$\hat{\mu_i} = \hat{\mathbf{S}}_i \bar{\mathbf{Z}}^T \mathbf{a}_{(\bullet),i}$$

$$\hat{\mathbf{S}_i}^{-1} = \mathbf{I} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

2.7.3 $p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ in the augmented probit regression model

In the augmented probit regression model, the posterior distribution of a_i is distributed according to a truncated normal distribution.

$$p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)$$

2.7.4 $p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right)$

In our model, response variables are not always observed and are treated as latent and sampled where appropriate. There are two factors influencing predictions of the response variable, $y_{m,i}$. There is an undirected model enforcing the aforementioned constraints and providing a prior and there is the probit regression.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi(\mathbf{Y})\,\delta\left(sign\left(a_{m,i}\right) = y_{m,i}\right)\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right) \tag{17}$$

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right) \tag{18}$$

Again, this conditional distribution can be evaluated through enumerations and normalization.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) = \frac{\psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)}{\sum_{y_{m,i}}\psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)} \tag{19}$$

3 Results

4 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[4] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.
[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

Supervised Topic Modeling in Clinical Text

Perotte A Li Y Pivovarov R Weiskopf N Wood F
Columbia University, New York, NY 10027, USA
{{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [4], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: International Challenge: Classifying Clinical Free Text Using Natural Language Processing (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Methods

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patients chief complaint, diagnostic findings, therapy administered, patients response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within

a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patients visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 Supervised Topic Models

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

2.4 Generative Model

Given the number of topics, K, the global prior over topic proportions, α', and the prior over topics, γ, the generative process for documents and responses is as follows:

1. For each topic:
   (a) Draw a distribution over words $\beta_k \sim Dir\left(u, \gamma\right)$
2. For each ICD9 Code:
   (a) Draw regression coefficient $\eta_i \mid \mu, \sigma \sim \mathcal{N}\left(\mu, \sigma\right)$
3. Draw a prior over topic proportions $m \mid \alpha' \sim Dir\left(u, \alpha'\right)$
4. For each document:
   (a) Draw topic proportions $\theta_d \mid \alpha \sim Dir\left(m, \alpha\right)$
   (b) For each word:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult\left(\theta_d\right)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult\left(\beta_{z_n}\right)$
   (c) For each ICD-9 code from the root of the hierarchy and recursively descending the tree:

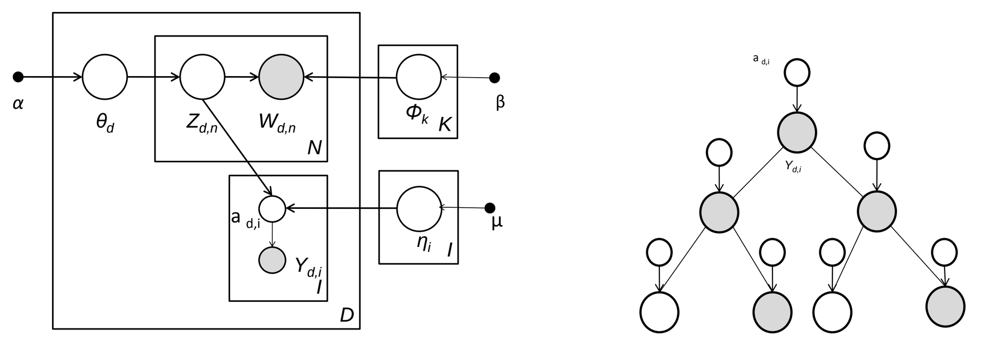


Figure 1: adapted sLDA model

i. Draw a response variable $y_i \mid \bar{z}, \eta_i, y_{parent} \sim \Phi(\eta_i^T \bar{z}, y_{parent})$ where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and Φ refers to a conditional probit model.

We will employ a data augmentation scheme with auxiliary variables a_i in the probit model where:

$$y_i \sim \begin{cases} 1, & a_i > 0 \; and \; y_{parent} = 1 \\ 0, & otherwise \end{cases} \tag{1}$$

$$a_i \sim \mathcal{N}\left(\eta_i^T \bar{z}, 1\right) \tag{2}$$

2.5 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation 4.

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler for the supervised topic model.

2.6 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{3}$$

$$\int_\theta \int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M} \left\{ p\left(\theta_m; \alpha\right) \prod_{n=1}^{N} \left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right] \prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right) \right\} \prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{4}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p\left(\theta_m; \alpha\right) \prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{5}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V}\Gamma\left(n^k_{(\bullet),v} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n^k_{(\bullet),v} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K}\Gamma\left(n^k_{m,(\bullet)} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n^k_{m,(\bullet)} + \alpha_k\right)} \tag{6}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right) \prod_{m=1}^{M} h\left(y\right) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V}\Gamma\left(n^k_{(\bullet),v} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n^k_{(\bullet),v} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K}\Gamma\left(n^k_{m,(\bullet)} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n^k_{m,(\bullet)} + \alpha_k\right)} \tag{7}$$

2.7 Gibbs Sampling

To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.7.1 $p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, m. The conditional probability with respect to this latent variable is proportional to the joint distribution up to a constant.

$$p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \propto p\left(z_{m,n}, \mathbf{z_{-(m,n)}}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \tag{8}$$

Due to the factorization of this model we can rewrite the joint distribution as the following:

$$\propto \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) p\left(y_{m,i} \mid z_{m,1:N}, \eta_i \xi\right)\right] p\left(z_{m,n}, \mathbf{z_{-(m,n)}}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \tag{9}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i=1}^{I}\left[p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\cdot),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\cdot),w_{m,n}} + \beta_{k,v}\right)} \tag{10}$$

Here, $n^{k,-(m,n)}_{m,v}$ represents the count of word v in document m assigned to topic k omitting the $(m,n)^{th}$ word count. For exponential family distributions, the normalization contant, $h(y_{m,i})$, does not depend on $z_{m,n}$.

$$\propto \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,v}\right)} \tag{11}$$

Given this expression, $p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right)$ can be sampled through enumeration as seen in Equation 13. In the case of probit regression, the expression for 11 evaluates to Equation 14. Equivalently we can parameterize the model with an auxiliary variable a_i, resulting in Equation 15.

$$p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) = \frac{\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,v}\right)}}{\sum_{k=1}^{K}\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,v}\right)}} \tag{12}$$

$$p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\Phi\left(\eta_i^T \bar{z}\right)^{y_{m,i}}\left(1 - \Phi\left(\eta_i^T \bar{z}\right)\right)^{1-y_{m,i}}\right]\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,v}\right)} \tag{13}$$

$$p\left(z_{m,n} \mid \mathbf{z_{-(m,n)}}, \mathbf{w}, \eta, \mathbf{a}, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\delta\left(sign\left(a_{m,i}\right) = 2y_{m,i} - 1\right)\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right)\right]\left(n^{k,-(m,n)}_{m,(\bullet)} + \alpha_k\right)\frac{n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n^{k,-(m,n)}_{(\bullet),w_{m,n}} + \beta_{k,v}\right)} \tag{14}$$

2.7.2 $p\left(\eta_i \mid \mathbf{z}, \mathbf{Y}, \mu\right)$ or $p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right)$in the augmented probit regression model

Given that η_i and $a_{m,i}$ are distributed normally, this posterior distribution is also normal. In the case for general exponential family distributions, η_i can remain a parameter without a prior, fit with maximimum likelihood in the usual fashion.

$$p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right) = \mathcal{N}\left(\eta_i \mid \hat{\mu_i}, \hat{\mathbf{S}}_i\right) \tag{15}$$

$$\hat{\mu_i} = \hat{\mathbf{S}}_i \bar{\mathbf{Z}}^T \mathbf{a}_{(\bullet),i}$$

$$\hat{\mathbf{S}}_i^{-1} = \mathbf{I} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

2.7.3 $p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ in the augmented probit regression model

In the augmented probit regression model, the posterior distribution of a_i is distributed according to a truncated normal distribution.

$$p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)$$

2.7.4 $p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right)$

In our model, response variables are not always observed and are treated as latent and sampled where appropriate. There are two factors influencing predictions of the response variable, $y_{m,i}$. There is an undirected model enforcing the aforementioned constraints and providing a prior and there is the probit regression.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi\left(\mathbf{Y}\right)\delta\left(sign\left(a_{m,i}\right) = y_{m,i}\right)\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right) \tag{16}$$

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right) \tag{17}$$

Again, this conditional distribution can be evaluated through enumerations and normalization.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) = \frac{\psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)}{\sum_{y_{m,i}}\psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)} \tag{18}$$

3 Results

4 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.

[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

Supervised Topic Modeling in Clinical Text

Perotte A **Li Y** **Pivovarov R** **Weiskopf N** **Wood F**
Columbia University, New York, NY 10027, USA
{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [?], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [?].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [? ? ? ?], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [?] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: International Challenge: Classifying Clinical Free Text Using Natural Language Processing (website). Most of the classification strategies included word matching and rule-based algorithms. [? ? ?]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [?]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Methods

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patients chief complaint, diagnostic findings, therapy administered, patients response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within

a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patients visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 Supervised Topic Models

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [?].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [?].

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [?]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [?]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

2.4 Generative Model

Given the number of topics, K, the global prior over topic proportions, α', and the prior over topics, γ, the generative process for documents and responses is as follows:

1. For each topic:
   (a) Draw a distribution over words $\beta_k \sim Dir(u, \gamma)$
2. For each ICD9 Code:
   (a) Draw regression coefficient $\eta_i \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma)$
3. Draw a prior over topic proportions $m \mid \alpha' \sim Dir(u, \alpha')$
4. For each document:
   (a) Draw topic proportions $\theta_d \mid \alpha \sim Dir(m, \alpha)$
   (b) For each word:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult(\beta_{z_n})$
   (c) For each ICD-9 code from the root of the hierarchy and recursively descending the tree:

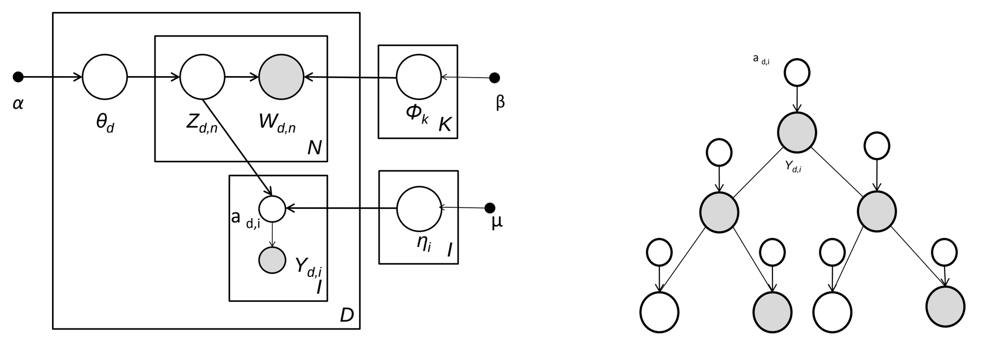


Figure 1: adapted sLDA model

i. Draw a response variable $y_i \mid \bar{z}, \eta_i, y_{parent} \sim \Phi(\eta_i^T \bar{z}, y_{parent})$ where $\bar{z} = N^{-1} \sum_{n=1}^{N} z_n$ and Φ refers to a conditional probit model.

We will employ a data augmentation scheme with auxiliary variables a_i in the probit model where:

$$y_i \sim \begin{cases} 1, & a_i > 0 \; and \; y_{parent} = 1 \\ 0, & otherwise \end{cases} \tag{1}$$

$$a_i \sim \mathcal{N}\left(\eta_i^T \bar{z}, 1\right) \tag{2}$$

2.5 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation 4.

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler for the supervised topic model.

2.6 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_{\theta} p\left(\theta \mid \alpha\right) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{3}$$

$$\int_{\theta}\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_{\theta}\int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right) \prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right] \prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\} \prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{4}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K} \int_{\theta} \prod_{m=1}^{M} p\left(\theta_m; \alpha\right) \prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{5}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{6}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right) \prod_{m=1}^{M} h\left(y\right) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{7}$$

2.7 Gibbs Sampling

To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.7.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, m. The conditional probability with respect to this latent variable is proportional to the joint distribution up to a constant.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \propto p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \tag{8}$$

Due to the factorization of this model we can rewrite the joint distribution as the following:

$$\propto \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \tag{9}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [?].

$$\propto \prod_{i=1}^{I}\left[p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{10}$$

Here, $n_{m,v}^{k,-(m,n)}$ represents the count of word v in document m assigned to topic k omitting the $(m,n)^{th}$ word count. For exponential family distributions, the normalization contant, $h(y_{m,i})$, does not depend on $z_{m,n}$.

$$\propto \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{11}$$

Given this expression, $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right)$ can be sampled through enumeration as seen in Equation 13. In the case of probit regression, the expression for ?? evaluates to Equation 14. Equivalently we can parameterize the model with an auxiliary variable a_i, resulting in Equation 15.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) = \frac{\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}}{\sum_{k=1}^{K} \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}} \tag{12}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\Phi\left(\eta_i^T \bar{z}\right)^{y_{m,i}}\left(1 - \Phi\left(\eta_i^T \bar{z}\right)\right)^{1-y_{m,i}}\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{13}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \mathbf{a}, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\delta\left(sign\left(a_{m,i}\right) = 2y_{m,i} - 1\right) \mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{14}$$

2.7.2 $p\left(\eta_i \mid \mathbf{z}, \mathbf{Y}, \mu\right)$ or $p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right)$ in the augmented probit regression model

Given that η_i and $a_{m,i}$ are distributed normally, this posterior distribution is also normal. In the case for general exponential family distributions, η_i can remain a parameter without a prior, fit with maximimum likelihood in the usual fashion.

$$p\left(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu\right) = \mathcal{N}\left(\eta_i \mid \hat{\mu}_i, \hat{\mathbf{S}}_i\right) \tag{15}$$

$$\hat{\mu}_i = \hat{\mathbf{S}}_i \bar{\mathbf{Z}}^T \mathbf{a}_{(\bullet),i}$$

$$\hat{\mathbf{S}}_i^{-1} = \mathbf{I} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

2.7.3 $p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ in the augmented probit regression model

In the augmented probit regression model, the posterior distribution of a_i is distributed according to a truncated normal distribution.

$$p\left(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)$$

2.7.4 $p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right)$

In our model, response variables are not always observed and are treated as latent and sampled where appropriate. There are two factors influencing predictions of the response variable, $y_{m,i}$. There is an undirected model enforcing the aforementioned constraints and providing a prior and there is the probit regression.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi\left(\mathbf{Y}\right) \delta\left(sign\left(a_{m,i}\right) = y_{m,i}\right) \mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right) \tag{16}$$

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) \propto \psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right) \tag{17}$$

Again, this conditional distribution can be evaluated through enumerations and normalization.

$$p\left(y_{m,i} \mid \eta, \mathbf{a}, \xi\right) = \frac{\psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)}{\sum_{y_{m,i}} \psi\left(\mathbf{Y}\right) trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)} \tag{18}$$

3 Results

4 Conclusion

Supervised Topic Modeling in Clinical Text

Perotte A **Li Y** **Pivovarov R** **Weiskopf N** **Wood F**
Columbia University, New York, NY 10027, USA
`{{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu`

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [?], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [?].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [? ? ? ?], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [?] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: International Challenge: Classifying Clinical Free Text Using Natural Language Processing (website). Most of the classification strategies included word matching and rule-based algorithms. [? ? ?]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [?]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Methods

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patients chief complaint, diagnostic findings, therapy administered, patients response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within

a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patients visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 Supervised Topic Models

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [?].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [?].

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [?]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [?]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

2.4 Generative Model

Given the number of topics, K, the global prior over topic proportions, α', and the prior over topics, γ, the generative process for documents and responses is as follows:

1. For each topic:
   (a) Draw a distribution over words $\beta_k \sim Dir(u, \gamma)$
2. For each ICD9 Code:
   (a) Draw regression coefficient $\eta_i \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma)$
3. Draw a prior over topic proportions $m \mid \alpha' \sim Dir(u, \alpha')$
4. For each document:
   (a) Draw topic proportions $\theta_d \mid \alpha \sim Dir(m, \alpha)$
   (b) For each word:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult(\beta_{z_n})$
   (c) For each ICD-9 code from the root of the hierarchy and recursively descending the tree:

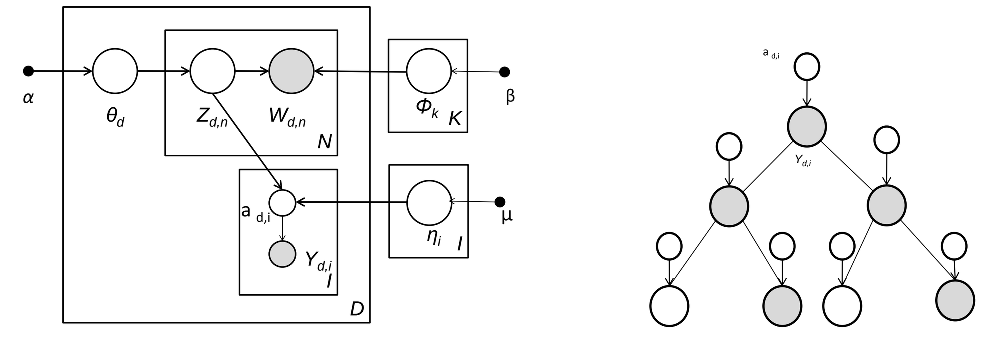


Figure 1: adapted sLDA model

i. Draw a response variable $y_i \mid \bar{z}, \eta_i, y_{parent} \sim \Phi(\eta_i^T \bar{z}, y_{parent})$ where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and Φ refers to a conditional probit model.

We will employ a data augmentation scheme with auxiliary variables a_i in the probit model where:

$$y_i \sim \begin{cases} 1, & a_i > 0 \; and \; y_{parent} = 1 \\ 0, & otherwise \end{cases} \tag{1}$$

$$a_i \sim \mathcal{N}(\eta_i^T \bar{z}, 1) \tag{2}$$

2.5 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation 4.

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler for the supervised topic model.

2.6 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right) d\theta} \tag{3}$$

$$\int_\theta \int_{\phi_{1:K}} p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi)\, d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M}\left\{ p(\theta_m; \alpha) \prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m)\, p(w_{m,n} \mid \phi_{z_{m,n}})\right] \prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\} \prod_{i=1}^{I} p(\eta_i; \mu)\, d\theta d\phi_{1:K} \tag{4}$$

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi) = \prod_{i=1}^{I}\left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M}\prod_{n=1}^{N} p(w_{m,n} \mid \phi_{z_{m,n}})\, d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p(\theta_m; \alpha) \prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)\, d\theta \tag{5}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{6}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1) \prod_{m=1}^{M} h(y) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{7}$$

2.7 Gibbs Sampling

To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.7.1 $p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, m. The conditional probability with respect to this latent variable is proportional to the joint distribution up to a constant.

$$p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi) \propto p(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu) \tag{8}$$

Due to the factorization of this model we can rewrite the joint distribution as the following:

$$\propto \prod_{i=1}^{I}\left[p(\eta_i; \mu)\, p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right] p(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu) \tag{9}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [?].

$$\propto \prod_{i=1}^{I}\left[p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{10}$$

Here, $n_{m,v}^{k,-(m,n)}$ represents the count of word v in document m assigned to topic k omitting the $(m,n)^{th}$ word count. For exponential family distributions, the normalization contant, $h(y_{m,i})$, does not depend on $z_{m,n}$.

$$\propto \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{11}$$

Given this expression, $p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu)$ can be sampled through enumeration as seen in Equation 13. In the case of probit regression, the expression for 11 evaluates to Equation 14. Equivalently we can parameterize the model with an auxiliary variable a_i, resulting in Equation 15.

$$p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu) = \frac{\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}}{\sum_{k=1}^{K} \exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}} \tag{12}$$

$$p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu) \propto \prod_{i=1}^{I}\left[\Phi\left(\eta_i^T \bar{z}\right)^{y_{m,i}}\left(1 - \Phi\left(\eta_i^T \bar{z}\right)\right)^{1-y_{m,i}}\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{13}$$

$$p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \mathbf{a}, \alpha, \beta, \mu) \propto \prod_{i=1}^{I}\left[\delta\left(sign(a_{m,i}) = 2y_{m,i} - 1\right) \mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{14}$$

2.7.2 $p(\eta_i \mid \mathbf{z}, \mathbf{Y}, \mu)$ or $p(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu)$ in the augmented probit regression model

Given that η_i and $a_{m,i}$ are distributed normally, this posterior distribution is also normal. In the case for general exponential family distributions, η_i can remain a parameter without a prior, fit with maximimum likelihood in the usual fashion.

$$p(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu) = \mathcal{N}\left(\eta_i \mid \hat{\mu_i}, \hat{\mathbf{S}}_i\right) \tag{15}$$

$$\hat{\mu_i} = \hat{\mathbf{S}}_i \bar{\mathbf{Z}}^T \mathbf{a}_{(\bullet),i}$$

$$\hat{\mathbf{S}}_i^{-1} = \mathbf{I} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

2.7.3 $p(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta)$ in the augmented probit regression model

In the augmented probit regression model, the posterior distribution of a_i is distributed according to a truncated normal distribution.

$$p(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta) = trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)$$

2.7.4 $p(y_{m,i} \mid \eta, \mathbf{a}, \xi)$

In our model, response variables are not always observed and are treated as latent and sampled where appropriate. There are two factors influencing predictions of the response variable, $y_{m,i}$. There is an undirected model enforcing the aforementioned constraints and providing a prior and there is the probit regression.

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) \propto \psi(\mathbf{Y})\, \delta\left(sign(a_{m,i}) = y_{m,i}\right) \mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right) \tag{16}$$

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) \propto \psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right) \tag{17}$$

Again, this conditional distribution can be evaluated through enumerations and normalization.

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) = \frac{\psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)}{\sum_{y_{m,i}} \psi(\mathbf{Y})\, trunc\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)} \tag{18}$$

3 Results

4 Conclusion

References

[] International classification of disease. `http://bioportal.bioontology.org/ontologies/35686`, May 2008.
[] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.
[] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

Supervised Topic Modeling in Clinical Text

Perotte A **Li Y** **Pivovarov R** **Weiskopf N** **Wood F**
Columbia University, New York, NY 10027, USA
{{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [4], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

3 Methods

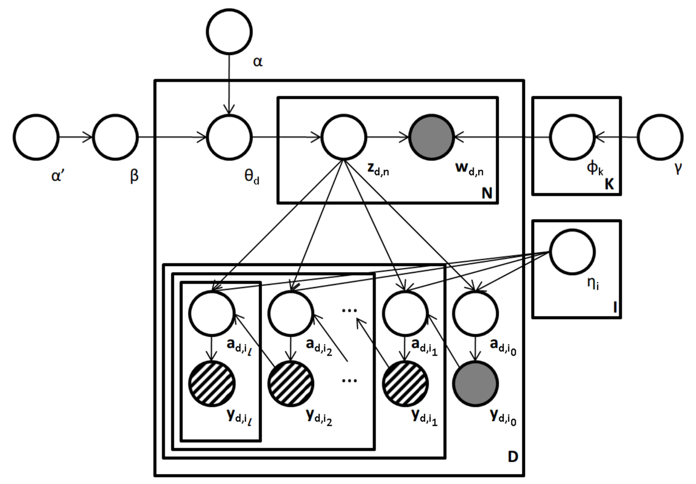


Figure 1: adapted sLDA model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable
$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$
where $\bar{z} = N^{-1} \sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$
         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right) \tag{1}$$
$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta, \phi, \eta, a, \alpha, \alpha', \beta, \gamma} \sum_{z} p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [10]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c} \in \mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i_{l,c} \in \mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T \eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} p\left(a_{i_{l,c}} \mid \eta_{i_{l,c}}\right) \left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right) \frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V} \left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation **??**, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i \left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}} \in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right), \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics **?**].

Posterior inference was performed using the "direct assignment" method of **?**].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

3.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

4 Results

5 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[4] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.
[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

5.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

- Benefits of combining human categorization information into "topic models"
-

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

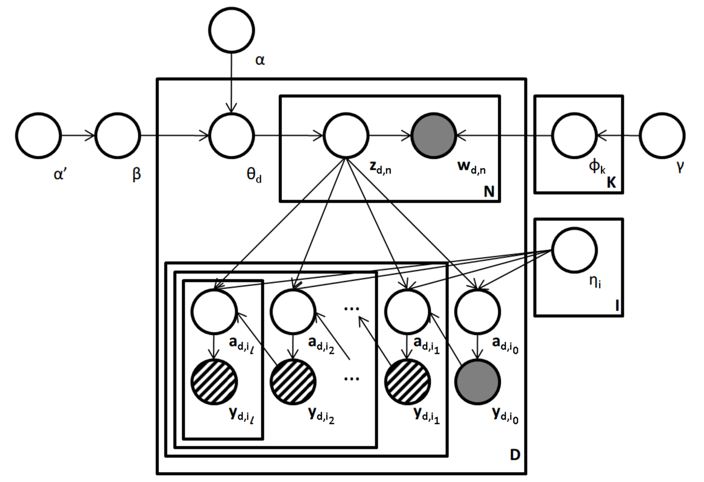


Figure 1: adapted sLDA model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [10]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} p\left(a_{i_{l,c}} \mid \eta_{i_{l,c}}\right)\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right), \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics **?**].

Posterior inference was performed using the "direct assignment" method of **?**].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable.

In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.

[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right)\prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p\left(\theta_m; \alpha\right)\prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right)\prod_{m=1}^{M} h\left(y\right)\exp\left\{\left(\eta_i^T\bar{z}\right) y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^k + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

- Benefits of combining human categorization information into "topic models"
-

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

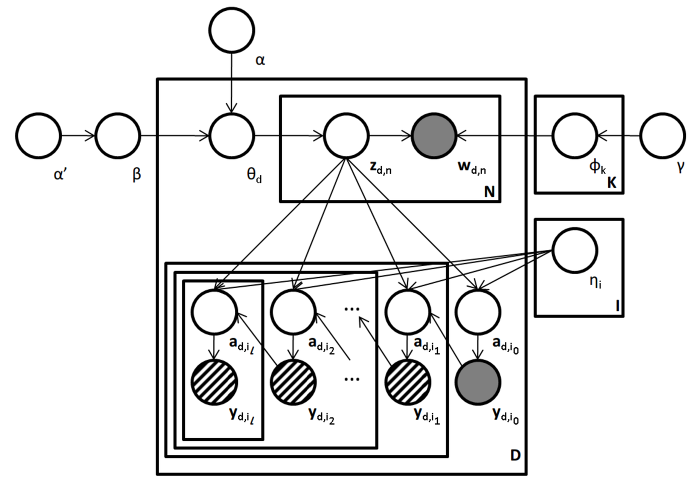


Figure 1: adapted sLDA model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1},\gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1},\sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1},\alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta,\alpha \sim Dir_K(\beta,\alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d},\beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable
$$a_{d,i_{l,c}} \sim \begin{cases}\mathcal{N}\left(\bar{z}^T\eta_{i_{l,c}},1\right), & y_{parent_{l,c}}=1\\ trunc\mathcal{N}^-\left(\bar{z}^T\eta_{i_{l,c}},1\right), & y_{parent_{l,c}}=-1\end{cases}$$
         where $\bar{z}=N^{-1}\sum_{n=1}^N z_n$ and $\mathcal{I}=\{i_0,i_1,...,i_{\mathcal{L}}\}$ and $i_l=\{i_{l,0},i_{l,1},...,i_{l,C_l}\}$
         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases}1, & a_{d,i_{l,c}}>0\\ -1, & otherwise\end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta,z_{1:N},\phi_{1:K},\eta_{i_{l,c}\in\mathcal{I}},a_{i_{l,c}\in\mathcal{I}},\beta,\alpha,\alpha',\gamma \mid w_{1:N},y_{i_{l,c}\in\mathcal{I}};\sigma,\lambda\right) \tag{1}$$

$$=\frac{p\left(\theta,z_{1:N},\phi_{1:K},\eta_{i_{l,c}\in\mathcal{I}},a_{i_{l,c}\in\mathcal{I}},\beta,\alpha,\alpha',\gamma,w_{1:N},y_{i_{l,c}\in\mathcal{I}};\sigma,\lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta,z_{1:N},\phi_{1:K},\eta_{i_{l,c}\in\mathcal{I}},a_{i_{l,c}\in\mathcal{I}},\beta,\alpha,\alpha',\gamma,w_{1:N},y_{i_{l,c}\in\mathcal{I}};\sigma,\lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [10]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})},\mathbf{a},\mathbf{w},\eta,\alpha,\beta,\gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})},\mathbf{a},\mathbf{w},\eta,\alpha,\beta,\gamma\right) \propto p\left(z_{d,n},\mathbf{z}_{-(\mathbf{d},\mathbf{n})},\mathbf{a},\mathbf{w},\eta,\alpha,\beta,\gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z},\eta_{i_{l,c}}\right) p\left(z_{d,n},\mathbf{z}_{-(\mathbf{d},\mathbf{n})},\mathbf{a},\mathbf{w},\eta,\alpha,\beta,\gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}}-a_{i_{l,c}}\right)^2}{2}\right\} p\left(a_{i_{l,c}} \mid \eta_{i_{l,c}}\right)\left(n_{d,(\cdot)}^{k,-(d,n)}+\alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)}+\gamma}{\sum_{v=1}^V\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)}+\gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})},\mathbf{a},\mathbf{w},\eta,\alpha,\beta,\gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D},\mathbf{a};\sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma=\{0.01,0.1,0.25,1,2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D},\mathbf{a};\sigma\right)=\mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i,\hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i=\hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1}+\bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1}=\mathbf{I}\sigma^{-1}+\bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z},\mathbf{Y},\eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z},\mathbf{Y},\eta\right)=\begin{cases}trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z},\mathbf{1},y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}}=1\\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z},\mathbf{1},y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}}=-1\end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}}>0$ and $y_{d,i_{l,c}}=1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}}<0 \mid y_{d,i_{l,c}}=-1\right)=1$ and $p\left(y_{d,i_{l,c}}=-1 \mid a_{d,i_{l,c}}<0\right)=1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}},y_{d,i_{l,c}} \mid \mathbf{z},\mathbf{Y}_{-(d,i_{l,c})},\eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a},\mathbf{y}_{-(l,c)}\right)p\left(a_{d,i_{l,c}} \mid \mathbf{z},\mathbf{Y},\eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a},\mathbf{y}_{-(l,c)}\right)=\delta\left(sign\left(a_{d,i_{l,c}}\right)=y_{i_{l,c}}\right)p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right)\prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}}=-1 \mid y_{parent_{l,c}}\right)=\begin{cases}1, & y_{parent_{l,c}}=-1\\ 0.5, & y_{parent_{l,c}}=1\end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}},y_{d,i_{l,c}} \mid \mathbf{z},\mathbf{Y}_{-(d,i_{l,c})},\eta\right)$$

$$=\begin{cases}\mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}},1\right)p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}}=1,\forall y_{i_{\hat{l},\hat{c}}}\in y_{children_{l,c}},y_{i_{\hat{l},\hat{c}}}=-1\\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}},1\right)\delta\left(y_{d,i_{l,c}}=-1\right), & y_{parent_{l,c}}=-1\\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}},1\right)\delta\left(y_{d,i_{l,c}}=1\right), & \exists y_{i_{\hat{l},\hat{c}}}\in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}}=1\\ 0 & otherwise\end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z},\alpha',\alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [17].

Posterior inference was performed using the "direct assignment" method of Teh et al. [16].

$$\beta \sim Dir\left(m_{(\cdot),1},m_{(\cdot),2},\ldots,m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k}=m \mid \mathbf{z},\mathbf{m}_{-(d,k)},\beta\right)=\frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k+n_{d,k}\right)}s\left(n_{d,k},m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n,m\right)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha;\lambda\right)$, $p\left(\alpha';\lambda\right)$, $p\left(\beta;\lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda=\{shape=2,scale=1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable.

In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[17] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta,z_{1:N} \mid w_{1:N},y_{1:I},\phi_{1:K},\eta_{1:I},\alpha,\beta,\mu,\xi\right)=\frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^N p\left(z_n \mid \theta\right)p\left(w_n \mid z_n,\phi_{1:K}\right)\right)\left(\prod_{k=1}^K p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^I p\left(y_i \mid z_{1:N},\eta_i,\xi\right)p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^K\left(\prod_{n=1}^N p\left(z_n \mid \theta\right)p\left(w_n \mid z_n,\phi_{1:K}\right)\right)\left(\prod_{k=1}^K p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^I p\left(y_i \mid z_{1:N},\eta_i,\xi\right)p\left(\eta_i \mid \mu\right)\right)d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y},\mathbf{w},\mathbf{z},\theta,\phi,\eta,\alpha,\beta,\mu,\xi\right)d\theta d\phi_{1:K}=\int_\theta\int_{\phi_{1:K}}\prod_{k=1}^K p\left(\phi_k;\beta\right)\prod_{m=1}^M\left\{p\left(\theta_m;\alpha\right)\prod_{n=1}^N\left[p\left(z_{m,n} \mid \theta_m\right)p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^I p\left(y_{m,i} \mid z_{m,1:N},\eta_i,\xi\right)\right\}\prod_{i=1}^I p\left(\eta_i;\mu\right)d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y},\mathbf{w},\mathbf{z},\eta,\alpha,\beta,\mu,\xi\right)=\prod_{i=1}^I\left[p\left(\eta_i;\mu\right)\prod_{m=1}^M p\left(y_{m,i} \mid z_{m,1:N},\eta_i,\xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^K p\left(\phi_k;\beta\right)\prod_{m=1}^M\prod_{n=1}^N p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)d\phi_{1:K}\int_\theta\prod_{m=1}^M p\left(\theta_m;\alpha\right)\prod_{n=1}^N p\left(z_{m,n} \mid \theta_m\right)d\theta \tag{15}$$

$$=\prod_{i=1}^I\left[p\left(\eta_i;\mu\right)\prod_{m=1}^M p\left(y_{m,i} \mid z_{m,1:N},\eta_i,\xi\right)\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\cdot),v}^k+\beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k+\beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\cdot)}^k+\alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k+\alpha_k\right)} \tag{16}$$

$$=\prod_{i=1}^I\left[\mathcal{N}\left(\eta_i \mid \mu,1\right)\prod_{m=1}^M h\left(y\right)\exp\left\{\left(\eta_i^T\bar{z}\right)y-A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\bullet),v}^k+\beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\bullet),v}^k+\beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\bullet)}^k+\alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\bullet)}^k+\alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

- Benefits of combining human categorization information into "topic models"
-

In this paper we describe the use of a topic model based on supervised latent dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

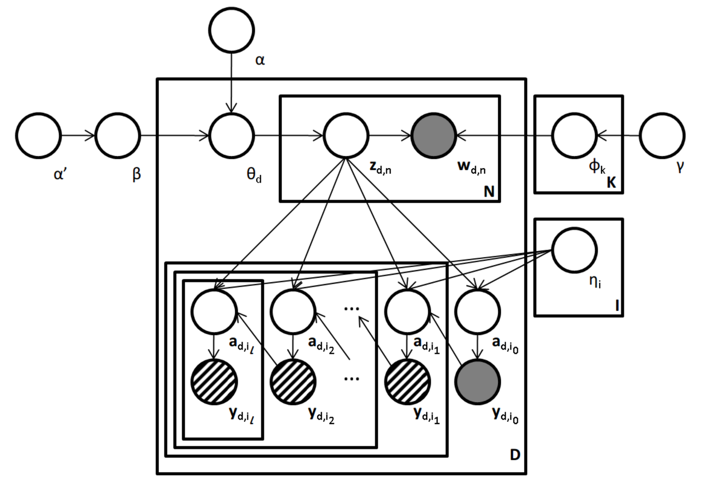


Figure 1: adapted sLDA model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable
$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$
where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$
         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma} \sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probablistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c} \in \mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c} \in \mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T \eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} \left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right) \frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appoporiately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}} \in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable.

In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{13}$$

$$\int_\theta \int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right) \prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right] \prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\} \prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M} \prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p\left(\theta_m; \alpha\right) \prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right) \prod_{m=1}^{M} h\left(y\right) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta\alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

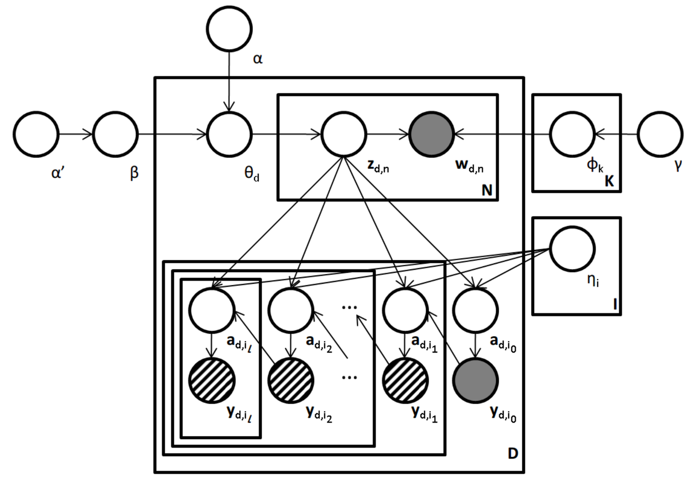


Figure 1: adapted sLDA model

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the structure of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\eta,a,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} p\left(a_{i_{l,c}} \mid \eta_{i_{l,c}}\right)\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-\{l,c\}}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-\{l,c\}}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,c}\in children_{l,c}} p\left(y_{i_{l,c}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,c}} \in y_{children_{l,c}}, y_{i_{l,c}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right), \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,c}} \in y_{children_{l,c}} \setminus y_{i_{l,c}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is

represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\left\{p(\theta_m; \alpha)\prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m)\, p(w_{m,n} \mid \phi_{z_{m,n}})\right]\prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\}\prod_{i=1}^{I} p(\eta_i; \mu)\, d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\prod_{n=1}^{N} p(w_{m,n} \mid \phi_{z_{m,n}})\, d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p(\theta_m; \alpha)\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)\, d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^{M} h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

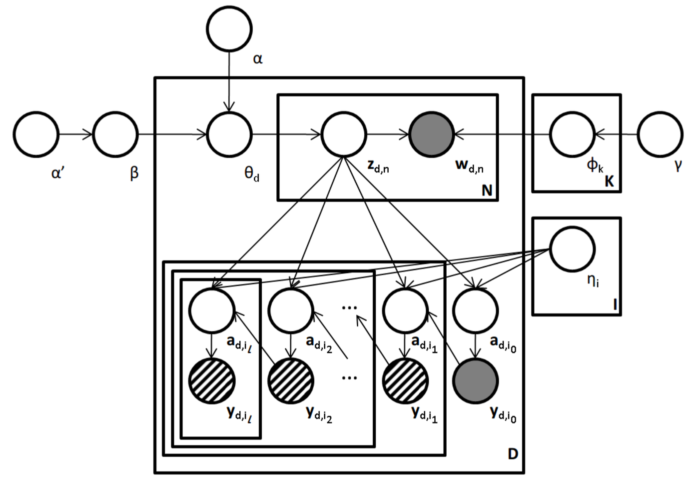


Figure 1: adapted sLDA model

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\eta,a,\beta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \bar{\mathbf{z}}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right)\prod_{i_{l,c}\in children_{l,c}} p\left(y_{i_{l,c}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,c}} \in y_{children_{l,c}}, y_{i_{l,c}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,c}} \in y_{children_{l,c}} \setminus y_{i_{l,c}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is

represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. `http://bioportal.bioontology.org/ontologies/35686`, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\left\{p(\theta_m; \alpha)\prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m)\, p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^{I} p(\eta_i; \mu)\, d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p(\theta_m; \alpha)\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)\, d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^{M} h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^k + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist surprisingly many sources of unstructured text data that have been partially or completely categorized by human editors. Examples include hierarchical directories of webpages [?], large hierarchically annotated product catalogs (e.g. [?] as available from [?]), manually annotated patient medical records, and many more. In this work we show how to combine these two sources of information in a single model that allows us to, amongst other things, automatically annotate and/or categorize new text documents, effectively inserting them into the category

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:

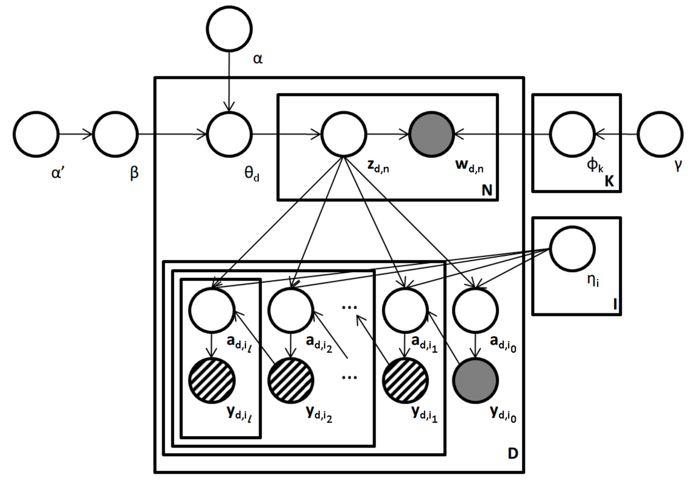


Figure 1: adapted sLDA model

A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1} \sum_{n=1}^N z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_\mathcal{L}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta, \phi, a, \eta, \alpha, \alpha', \beta, \gamma} \sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c} \in \mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c} \in \mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T \eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} \left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right) \frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^V \left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i \left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{i_{l,c}}}\right) \prod_{i_{l,\xi} \in children_{i_{l,c}}} p\left(y_{i_{l,\xi}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_l}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,\xi}} \in y_{children_{l,c}}, y_{i_{l,\xi}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,\xi}} \in y_{children_{l,c}} \setminus y_{i_{l,\xi}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha) \sum_{k=1}^K \left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta \int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^K p(\phi_k; \beta) \prod_{m=1}^M \left\{ p(\theta_m; \alpha) \prod_{n=1}^N \left[p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \phi_{z_{m,n}})\right] \prod_{i=1}^I p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right\} \prod_{i=1}^I p(\eta_i; \mu) \, d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^I \left[p(\eta_i; \mu) \prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right] \int_{\phi_{1:K}} \prod_{k=1}^K p(\phi_k; \beta) \prod_{m=1}^M \prod_{n=1}^N p(w_{m,n} \mid \phi_{z_{m,n}}) \, d\phi_{1:K} \int_\theta \prod_{m=1}^M p(\theta_m; \alpha) \prod_{n=1}^N p(z_{m,n} \mid \theta_m) \, d\theta \tag{15}$$

$$= \prod_{i=1}^I \left[p(\eta_i; \mu) \prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right] \prod_{k=1}^K \frac{\Gamma\left(\sum_{v=1}^V \beta_v\right)}{\prod_{v=1}^V \Gamma(\beta_v)} \frac{\prod_{v=1}^V \Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k + \beta_v\right)} \prod_{m=1}^M \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^I \left[\mathcal{N}(\eta_i \mid \mu, 1) \prod_{m=1}^M h(y) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\} \right] \prod_{k=1}^K \frac{\Gamma\left(\sum_{v=1}^V \beta_v\right)}{\prod_{v=1}^V \Gamma(\beta_v)} \frac{\prod_{v=1}^V \Gamma\left(n_{(\bullet),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\bullet),v}^k + \beta_v\right)} \prod_{m=1}^M \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma\left(n_{m,(\bullet)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\bullet)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist surprisingly many sources of unstructured text data that have been partially or completely categorized by human editors. Examples include hierarchical directories of webpages [?], large hierarchically annotated product catalogs (e.g. [?] as available from [?]), manually annotated patient medical records, and many more. In this work we show how to combine these two sources of information in a single model that allows us to, amongst other things, automatically annotate and/or categorize new text documents, effectively inserting them into the category

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:

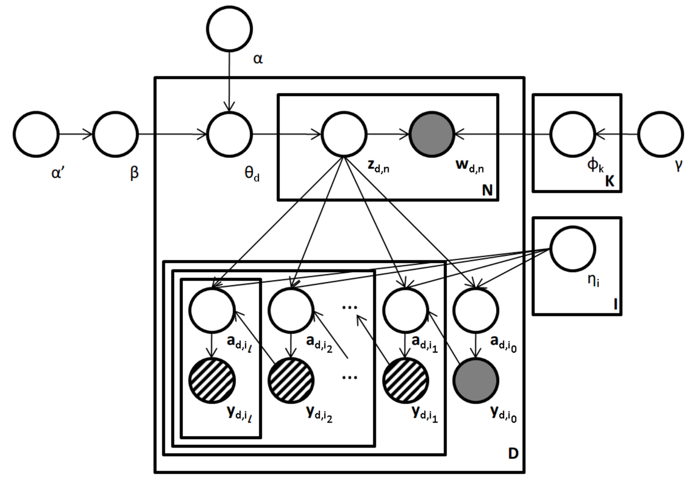


Figure 1: adapted sLDA model

A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma} \sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right)\prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p\left(\theta_m; \alpha\right)\prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right)\prod_{m=1}^{M} h\left(y\right)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**

Columbia University, New York, NY 10027, USA

{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist surprisingly many sources of unstructured text data that have been partially or completely categorized by human editors. Examples include hierarchical directories of webpages [?], large hierarchically annotated product catalogs (e.g. [?] as available from [?]), manually annotated patient medical records, and many more. In this work we show how to combine these two sources of information in a single model that allows us to, amongst other things, automatically annotate and/or categorize new text documents, effectively inserting them into the category

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:

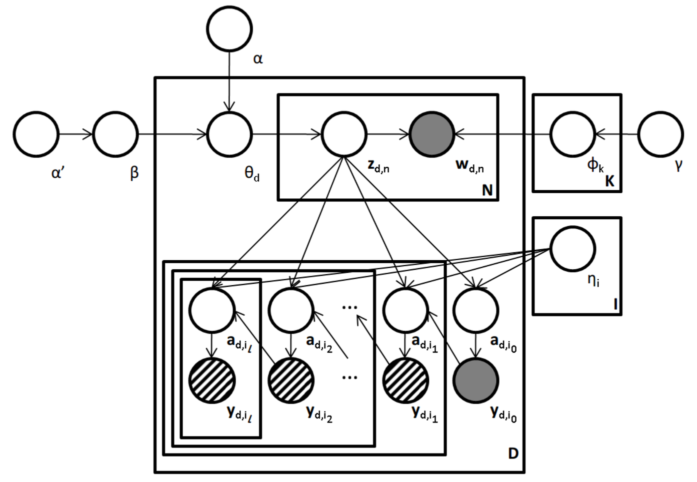


Figure 1: adapted sLDA model

A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}(\bar{z}^T \eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-(\bar{z}^T \eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda) \tag{1}$$

$$= \frac{p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda)}{\int_{\theta,\phi,\eta,a,\alpha,\alpha',\beta,\gamma} \sum_z p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto p(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}) \, p(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{(\bar{z}^T \eta_{i_{l,c}} - a_{i_{l,c}})^2}{2}\right\} \left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right) \frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$ can be sampled through enumeration.

3.2.2 $p(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma) = \mathcal{N}(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i \left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

3.2.3 $p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta)$ and $p(y_{m,i} \mid \mathbf{a})$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta) = \begin{cases} trunc\mathcal{N}^+(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T \bar{z}, \mathbf{1}, y_{d,i_{l,c}}) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1) = 1$ and $p(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta) \propto p(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}) \, p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta) \tag{7}$$

$$p(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}) = \delta(sign(a_{d,i_{l,c}}) = y_{i_{l,c}}) \, p(y_{i_{l,c}} \mid y_{parent_{l,c}}) \prod_{i_{l,c}\in children_{l,c}} p(y_{i_{l,c}} \mid y_{i_l}) \tag{8}$$

$$p(y_{i_{l,c}} = -1 \mid y_{parent_l}) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta)$$

$$= \begin{cases} \mathcal{N}(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1) \, p(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}), & y_{parent_{l,c}} = 1, \forall y_{l,c} \in y_{children_{l,c}}, y_{l,c} = -1 \\ trunc\mathcal{N}^-(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1) \, \delta(y_{d,i_{l,c}} = -1), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1) \, \delta(y_{d,i_{l,c}} = 1), & \exists y_{l,c} \in y_{children_{l,c}} \setminus y_{l,c} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p(\beta \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m) (\alpha\beta_k)^m \tag{12}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta \int_{\phi_{1:K}} p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi) \, d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M} \left\{ p(\theta_m; \alpha) \prod_{n=1}^{N} \left[p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \phi_{z_{m,n}})\right] \prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right\} \prod_{i=1}^{I} p(\eta_i; \mu) \, d\theta d\phi_{1:K} \tag{14}$$

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi) = \prod_{i=1}^{I} \left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M} \prod_{n=1}^{N} p(w_{m,n} \mid \phi_{z_{m,n}}) \, d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p(\theta_m; \alpha) \prod_{n=1}^{N} p(z_{m,n} \mid \theta_m) \, d\theta \tag{15}$$

$$= \prod_{i=1}^{I} \left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi) \right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I} \left[\mathcal{N}(\eta_i \mid \mu, 1) \prod_{m=1}^{M} h(y) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\} \right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [**?**], product descriptions and catalogs (e.g. [**?**] as available from [**?**]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records and the International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned to them[]). In this work we show how to combine these two sources of information using a single model that allows one, among other things, to automatically categorize new text documents, suggest labels that might be inaccurate, and compute improved similarities between documents for information retrieval purposes. The models and techniques that we develop in this paper are applicable in other domains as well, for instance, unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bags of features image representations).

In this work we extent supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) augmented with per document "supervision" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data generated about a document; for instance its quality or relevance (e.g. online reviews), marks given to written work (e.g. graded essays), or the number of times a web document is linked. These labels are usually modeled as having been generated conditioned on the mix of topics found in a document. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. As one concrete example consider web retailers. They often have both a browse-able hierarchy and free-text descriptions of all products they sell. The situation of each product in the product hierarchy (often multiply situated) can be seen as a form of multiple labeling and as a similar products based on the similarity of their textual description is an u. An equivalent challenge, particularly for larger retailers, is to situate the merchandise in as many categories as possible.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$

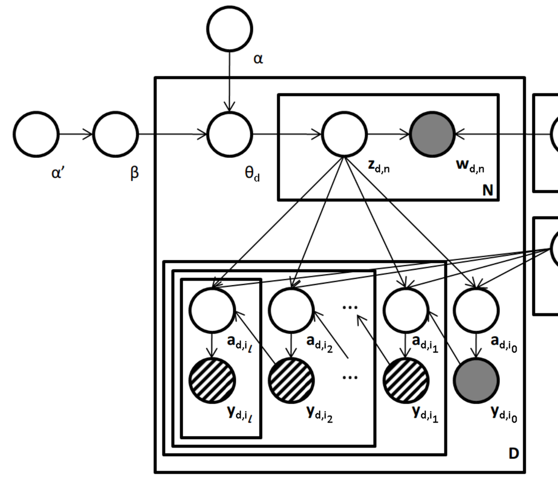


Figure 1: adapted sLDA model

4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}(\bar{z}^T \eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-(\bar{z}^T \eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^N z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda) \tag{1}$$

$$= \frac{p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda)}{\int_{\theta,\phi,\eta,a,\alpha,\alpha',\beta,\gamma}\sum_z p(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the distribution of its markov blanket up to a constant.

$$p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto p(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}})\, p(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}})^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^V\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$ can be sampled through enumeration.

3.2.2 $p(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma) = \mathcal{N}(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \mathbf{Z}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \mathbf{Z}^T\mathbf{Z}$$

3.2.3 $p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta)$ and $p(y_{m,i} \mid \mathbf{a})$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta) = \begin{cases} trunc\mathcal{N}^+(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1) = 1$ and $p(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta) \propto p(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{Y}_{-(l,c)})\, p(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta) \tag{7}$$

$$p(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{Y}_{-(l,c)}) = \delta(sign(a_{d,i_{l,c}}) = y_{i_{l,c}})\, p(y_{i_{l,c}} \mid y_{parent_{l,c}}) \prod_{i_{l,t}\in children_{l,c}} p(y_{i_{l,t}} \mid y_{i_{l,c}}) \tag{8}$$

$$p(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta)$$

$$= \begin{cases} \mathcal{N}(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1)\, p(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}), & y_{parent_{l,c}} = 1, \forall y_{i_{l,t}} \in y_{children_{l,c}}, y_{i_{l,t}} = -1 \\ trunc\mathcal{N}^-(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1)\, \delta(y_{d,i_{l,c}} = -1), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1)\, \delta(y_{d,i_{l,c}} = 1), & \exists y_{i_{l,t}} \in y_{children_{l,c}} \setminus y_{i_{l,t}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p(\beta \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}) \tag{11}$$

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m)(\alpha\beta_k)^m \tag{12}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.
[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

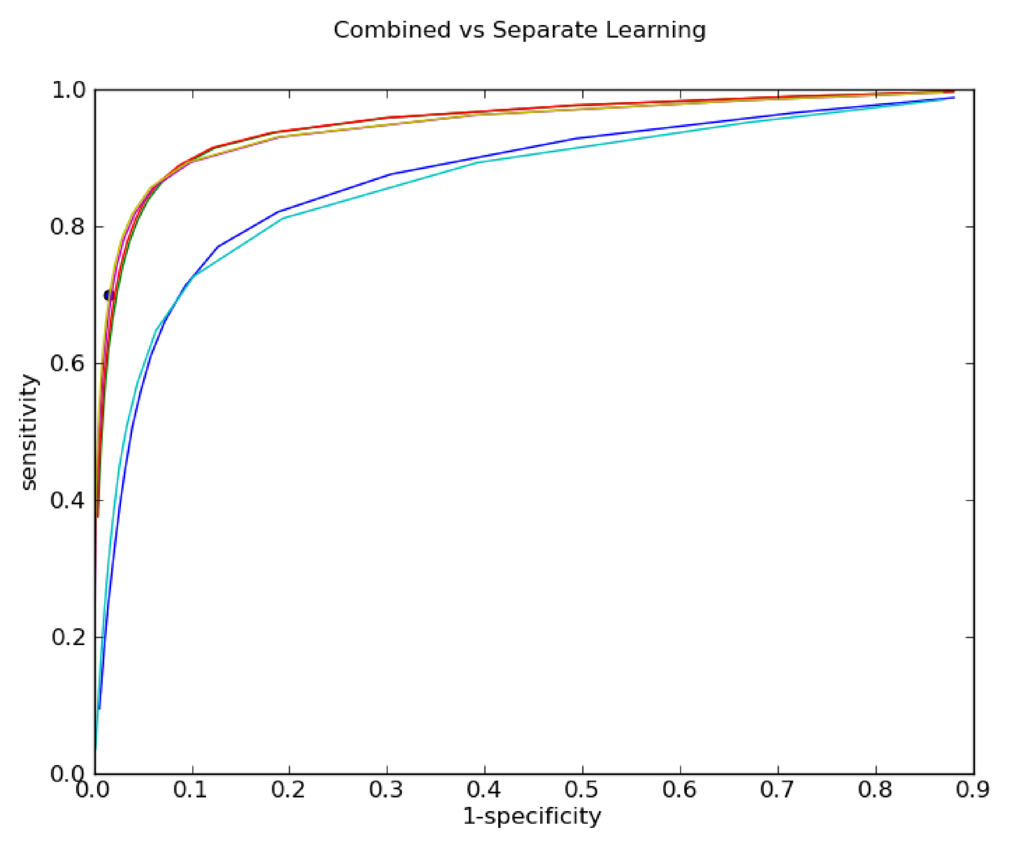


Figure 2: default

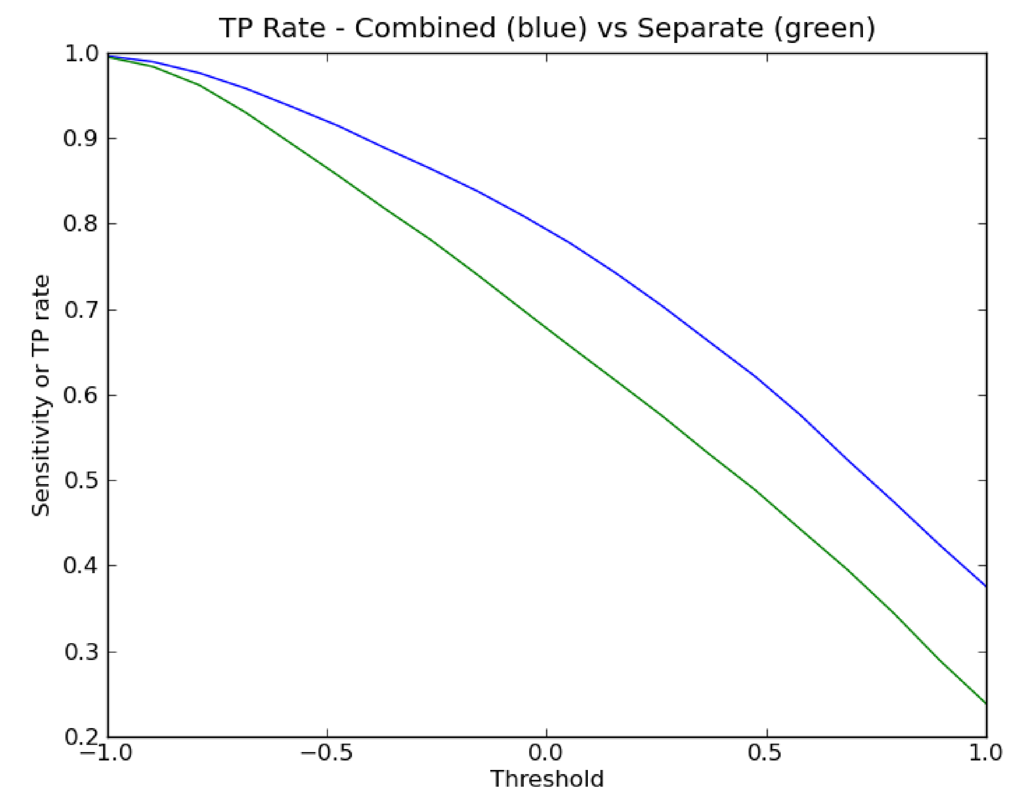


Figure 3: default

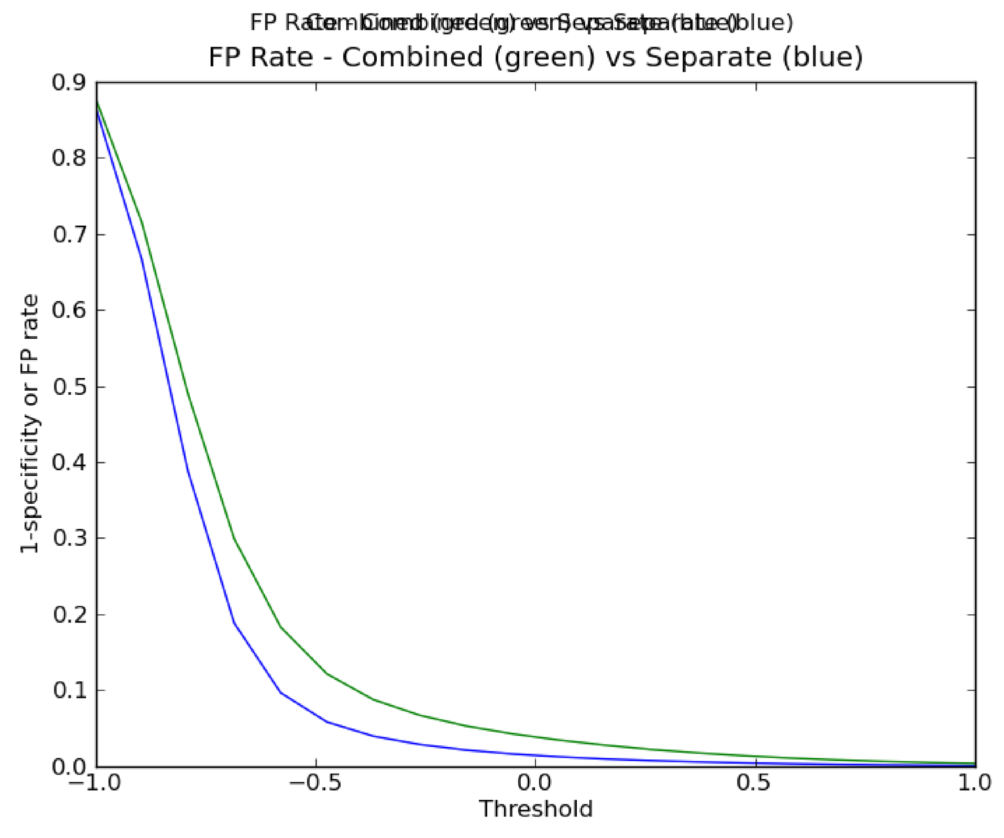


Figure 4: default

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist surprisingly many sources of unstructured text data that have been partially or completely categorized by human editors. Examples include hierarchical directories of webpages [?], large hierarchically annotated product catalogs (e.g. [?] as available from [?]), manually annotated patient medical records, and many more. In this work we show how to combine these two sources of information in a single model that allows us to, amongst other things, automatically annotate and/or categorize new text documents, effectively inserting them into the category

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly, viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:

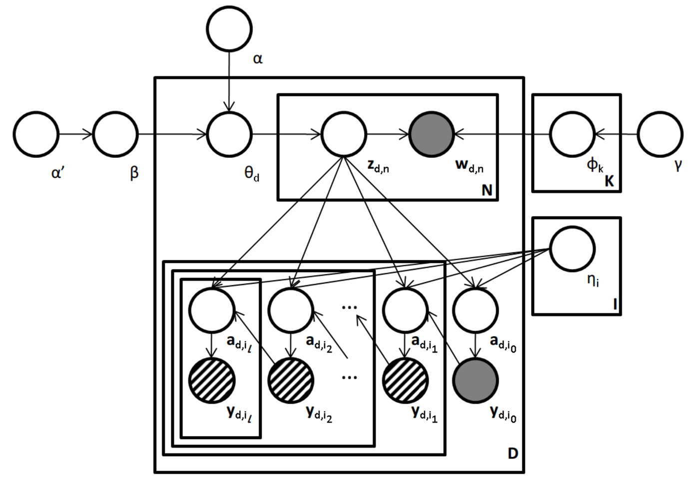


Figure 1: adapted sLDA model

A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\eta,a,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,c}\in children_{l,c}} p\left(y_{i_{l,c}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,c}} \in y_{children_{l,c}}, y_{i_{l,c}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,c}} \in y_{children_{l,c}} \setminus y_{i_{l,c}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Conclusion

References

[1] International classification of disease. `http://bioportal.bioontology.org/ontologies/35686`, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M}\left\{p(\theta_m; \alpha)\prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \phi_{z_{m,n}})\right]\prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\}\prod_{i=1}^{I} p(\eta_i; \mu) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\prod_{n=1}^{N} p(w_{m,n} \mid \phi_{z_{m,n}}) d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p(\theta_m; \alpha)\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m) d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^{M} h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [?], product descriptions and catalogs (e.g. [?] as available from [?]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records and the International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned to them[]). In this work we show how to combine these two sources of information using a single model that allows one, among other things, to automatically categorize new text documents, suggest labels that might be inaccurate, and compute improved similarities between documents for information retrieval purposes. The models and techniques that we develop in this paper are applicable in other domains as well, for instance, unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bags of features image representations).

In this work we extent supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) augmented with per document "supervision" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data generated about a document; for instance its quality or relevance (e.g. online reviews), marks given to written work (e.g. graded essays), or the number of times a web document is linked. These labels are usually modeled as having been generated conditioned on the mix of topics found in a document. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. As one concrete example consider web retailers. They often have both a browse-able hierarchy and free-text descriptions of all products they sell. The situation of each product in the product hierarchy (often multiply situated) can be seen as a form of multiple labeling and as a similar products based on the similarity of their textual description is an u. An equivalent challenge, particularly for larger retailers, is to situate the merchandise in as many categories as possible.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$

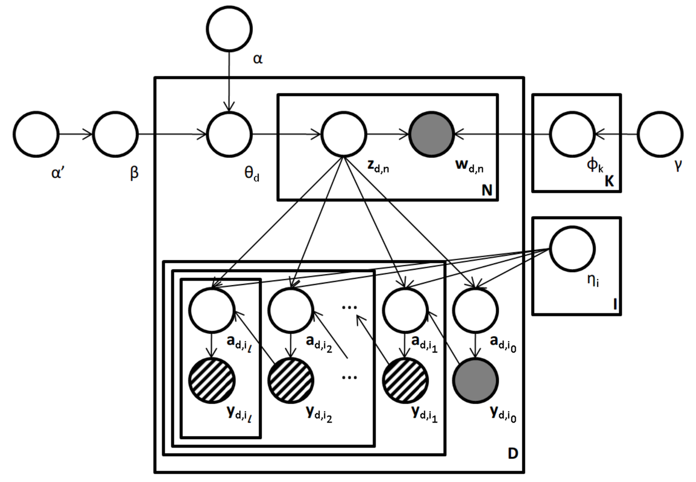


Figure 1: adapted sLDA model

4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1} \sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta, \phi, \eta, a, \alpha, \alpha', \beta, \gamma} \sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c} \in \mathcal{I}}, a_{i_{l,c} \in \mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c} \in \mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(m,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c} \in \mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c} \in \mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T \eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\} \left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right) \frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V} \left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k emitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i \left(-\mathbf{1}\sigma^{-1} + \mathbf{Z}^T \mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \mathbf{Z}^T\mathbf{Z}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T \bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T \bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{Y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{Y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,c} \in children_{l,c}} p\left(y_{i_{l,c}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,c}} \in y_{children_{l,c}}, y_{i_{l,c}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T \eta_{i_{l,c}}, 1\right) \delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,c}} \in y_{children_{l,c}} \setminus y_{i_{l,c}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

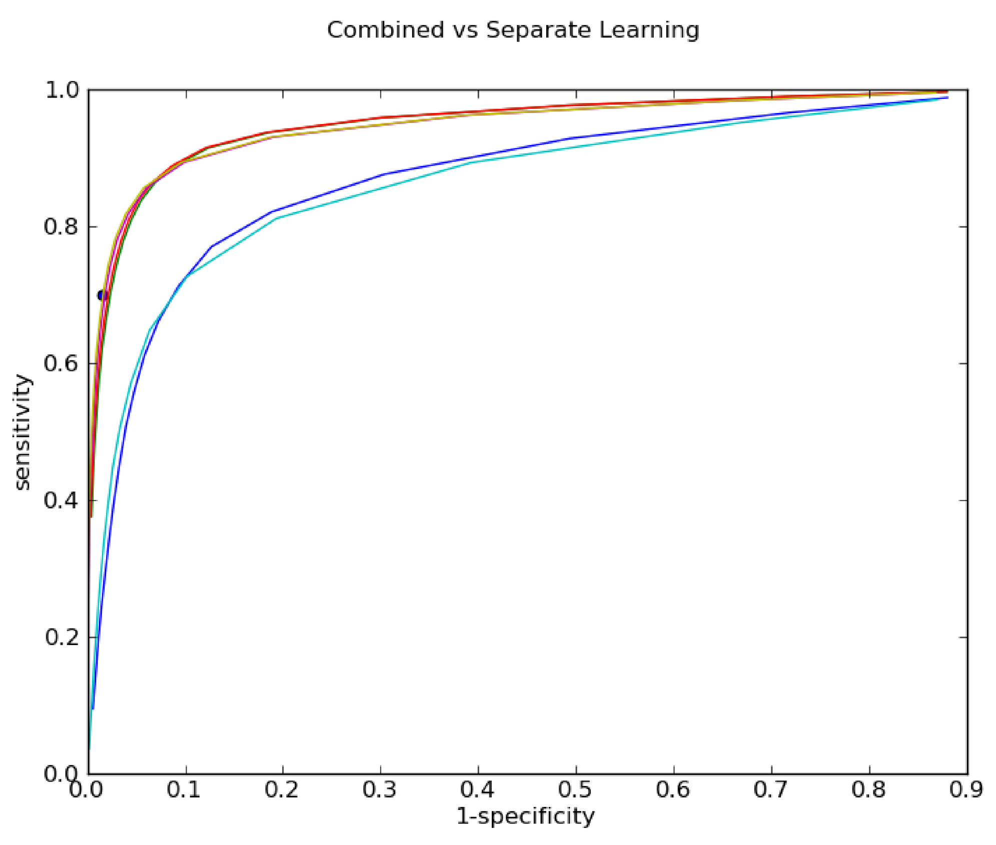


Figure 2: default

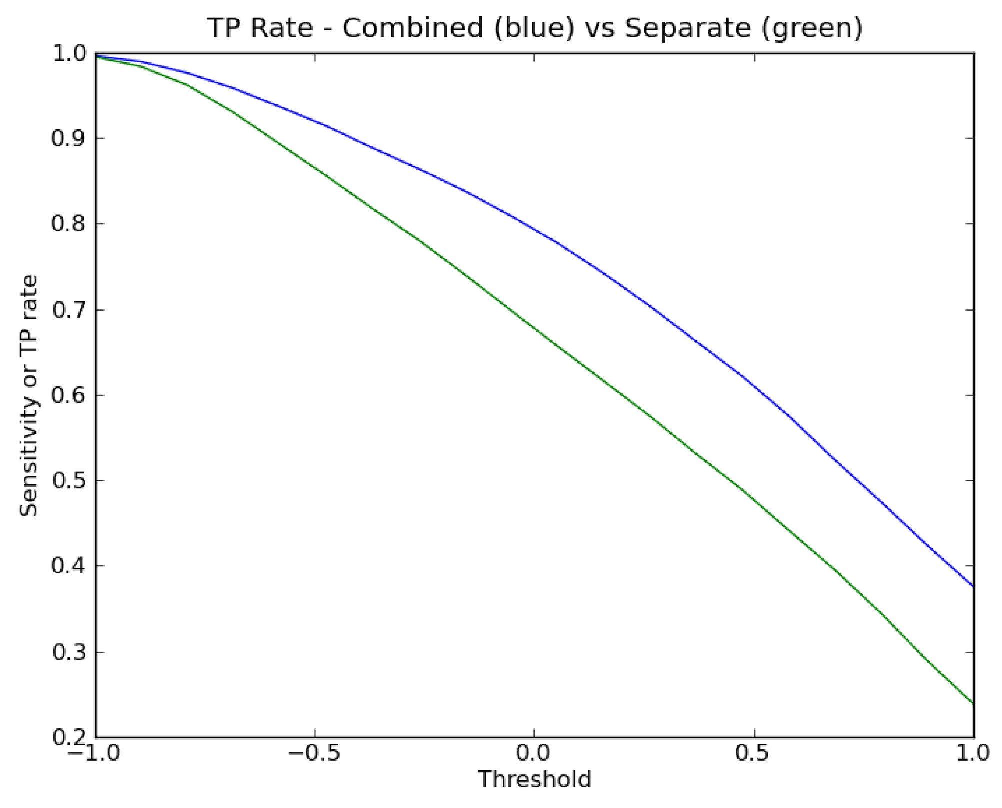


Figure 3: default

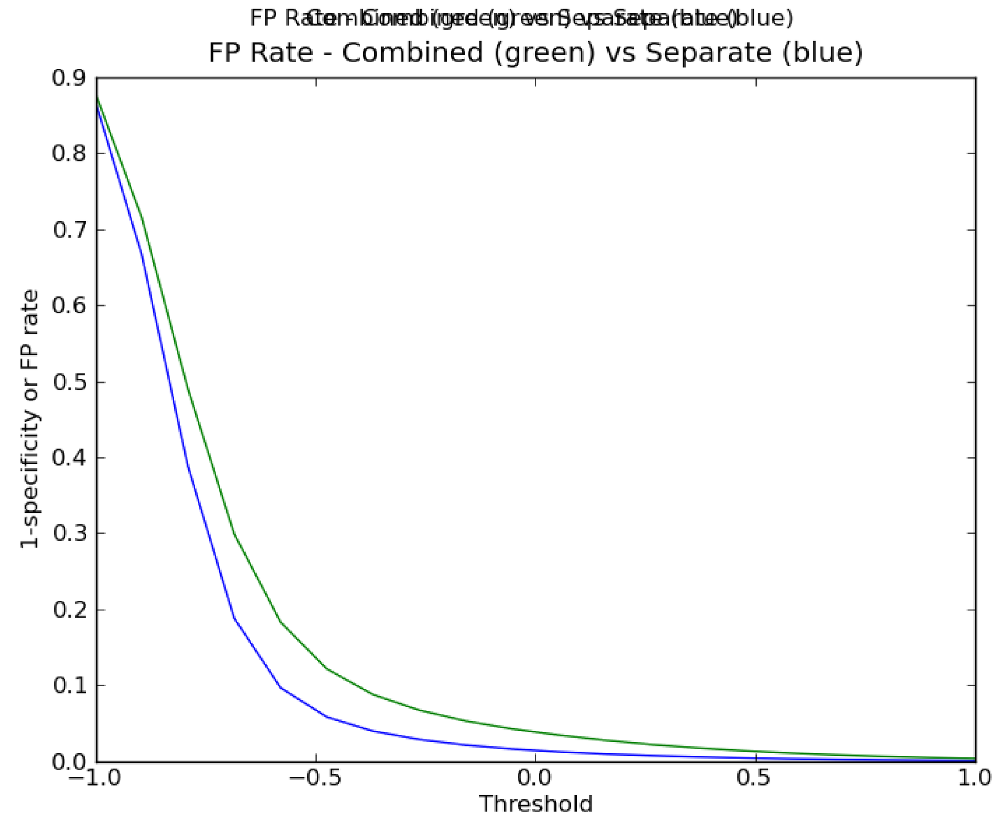


Figure 4: default

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta \int_{\phi_{1:K}} p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi)\, d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M}\left\{p(\theta_m; \alpha) \prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right] \prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\} \prod_{i=1}^{I} p(\eta_i; \mu)\, d\theta d\phi_{1:K} \tag{14}$$

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi) = \prod_{i=1}^{I}\left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{m=1}^{M} \prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p(\theta_m; \alpha) \prod_{n=1}^{N} p(z_{m,n} \mid \theta_m)\, d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu) \prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1) \prod_{m=1}^{M} h(y) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

The benefits of supervision in topic modeling

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [**?**], product descriptions and catalogs (e.g. [**?**] as available from [**?**]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records and the International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned to them[]). In this work we show how to combine these two sources of information using a single model that allows one, among other things, to automatically categorize new text documents, suggest labels that might be inaccurate, and compute improved similarities between documents for information retrieval purposes. The models and techniques that we develop in this paper are applicable in other domains as well, for instance, unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bags of features image representations).

In this work we extent supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) augmented with per document "supervision" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data generated about a document; for instance its quality or relevance (e.g. online reviews), marks given to written work (e.g. graded essays), or the number of times a web document is linked. These labels are usually modeled as having been generated conditioned on the mix of topics found in a document. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. As one concrete example consider web retailers. They often have both a browse-able hierarchy and free-text descriptions of all products they sell. The situation of each product in the product hierarchy (often multiply situated) can be seen as a form of multiple labeling and as a similar products based on the similarity of their textual description is an u. An equivalent challenge, particularly for larger retailers, is to situate the merchandise in as many categories as possible.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [14].

- Benefits of combining human categorization information into "topic models"
- LDA solved free text
- supervised LDA improves LDA (extra info) and allows new inference (predict links, etc.)
- amazon, freshdirect, netflix, dmoz, pandora (music genome)

2 Background

3 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$

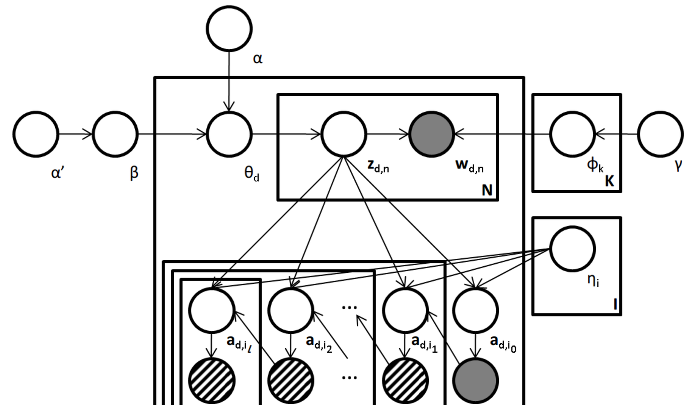


Figure 1: adapted sLDA model

4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.2.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.3 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Conclusion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.
[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.
[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

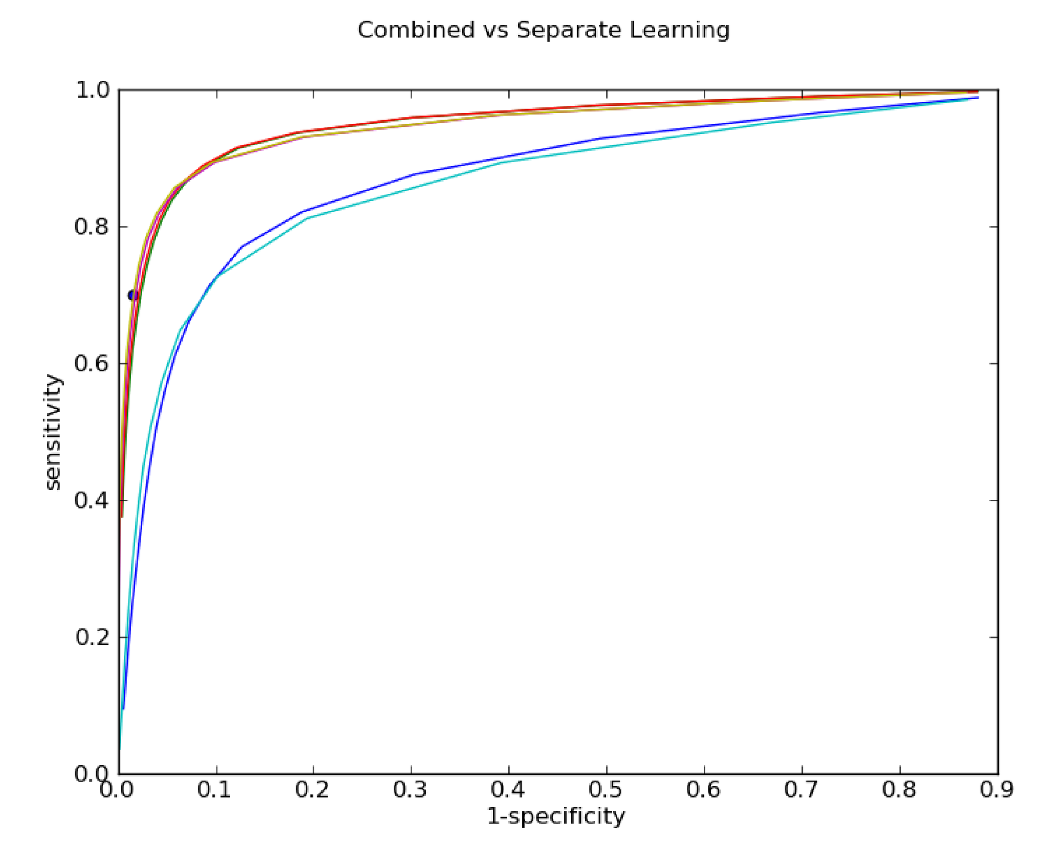


Figure 2: default

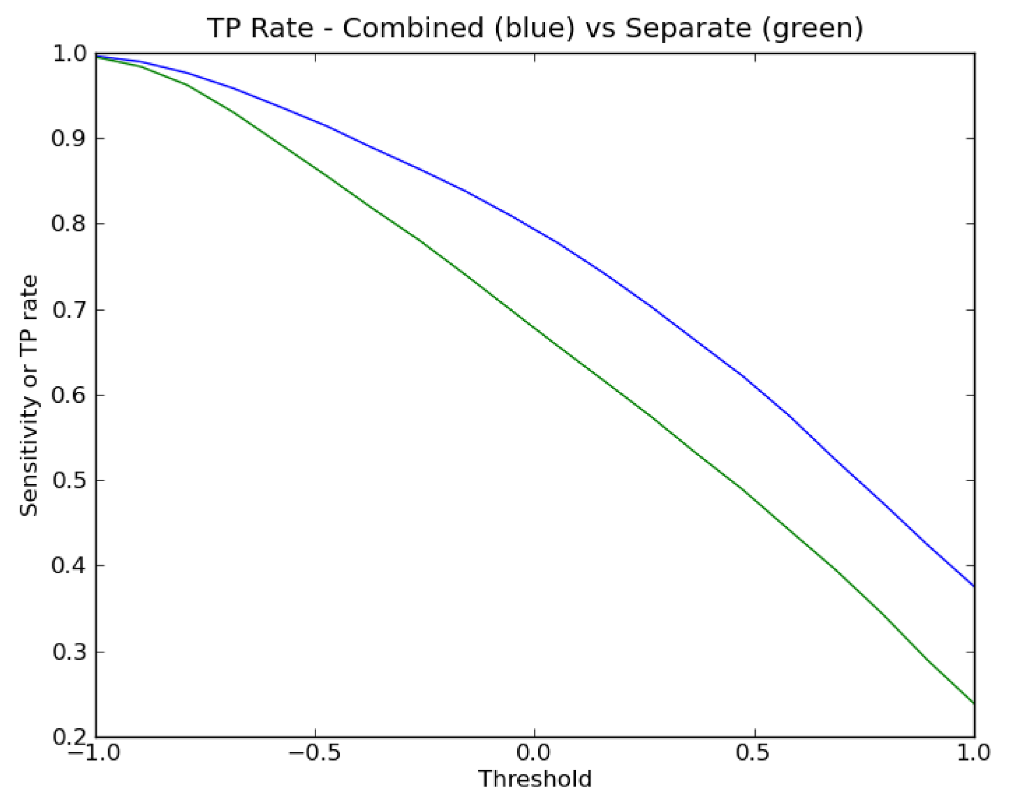


Figure 3: default

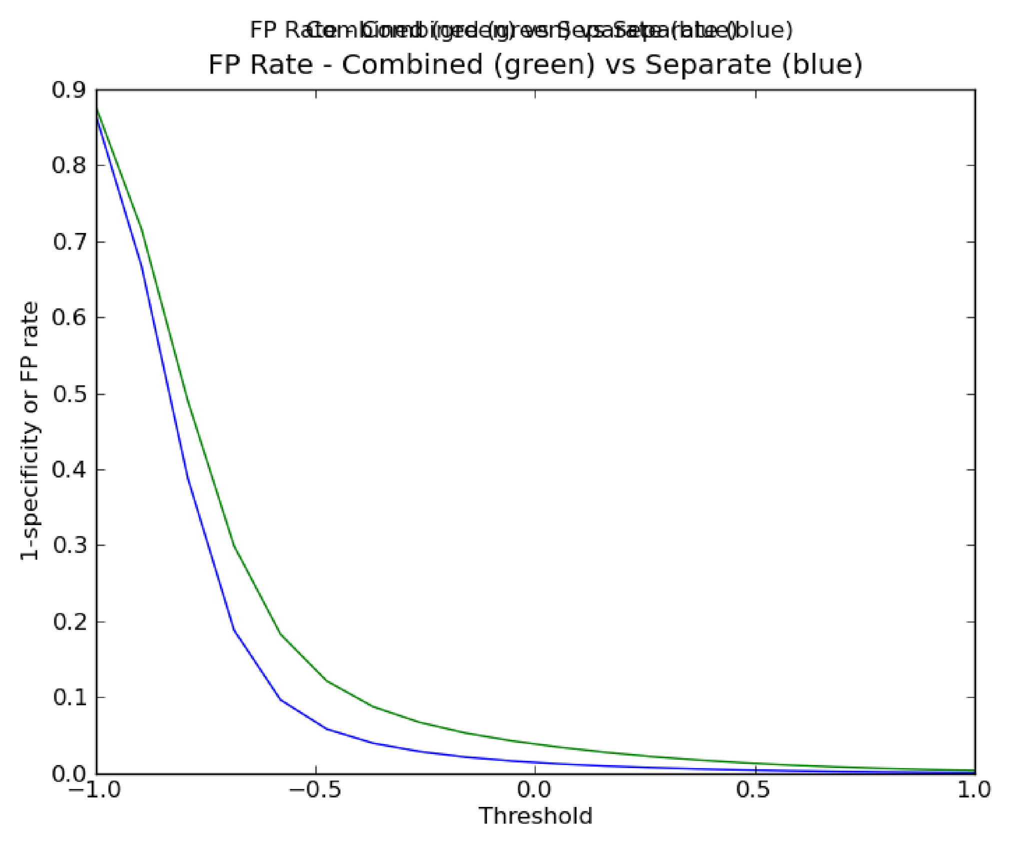


Figure 4: default

[illegible] (13)

[illegible] (14)

[illegible] (15)

[illegible] (16)

[illegible] (17)

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA) in the same manner as was done in supervised LDA (SLDA) prior art. We find that the additional supervision signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in medical document labeling and product categorization tasks. Additionally, held-out likelihood

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [?], product descriptions and catalogs (e.g. [?] as available from [?]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [3] augmented with per document "supervision," often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section ?? we introduce hierarchically supervised LDA (HSLDA), in Section ?? we review related work, and in Section ?? we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{l_{i,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$

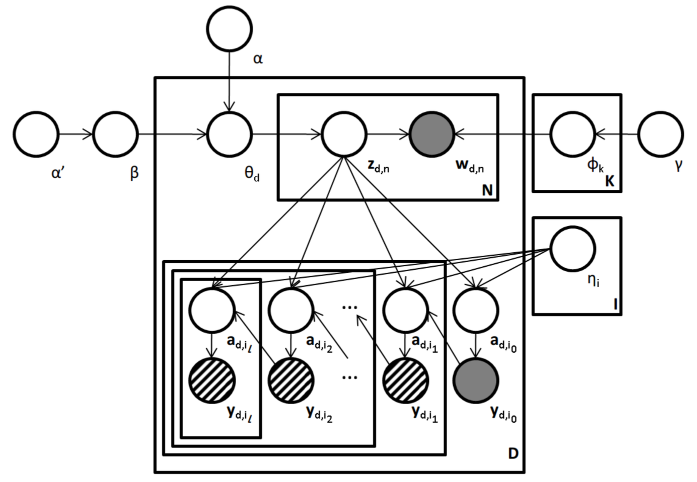


Figure 1: adapted sLDA model

      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^N z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

2.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\eta,a,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

2.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^V\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

2.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, a_{i_{l,c}}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

2.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,l} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation ??.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,j}\in children_{l,c}} p\left(y_{i_{l,j}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,j}} \in y_{children_{l,c}}, y_{i_{l,j}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,j}} \in y_{children_{l,c}} \setminus y_{i_{l,j}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

2.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

2.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

2.3 Prediction

3 Experiments

3.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

3.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

3.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

3.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4 Results

5 Discussion

References

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[14] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

5.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

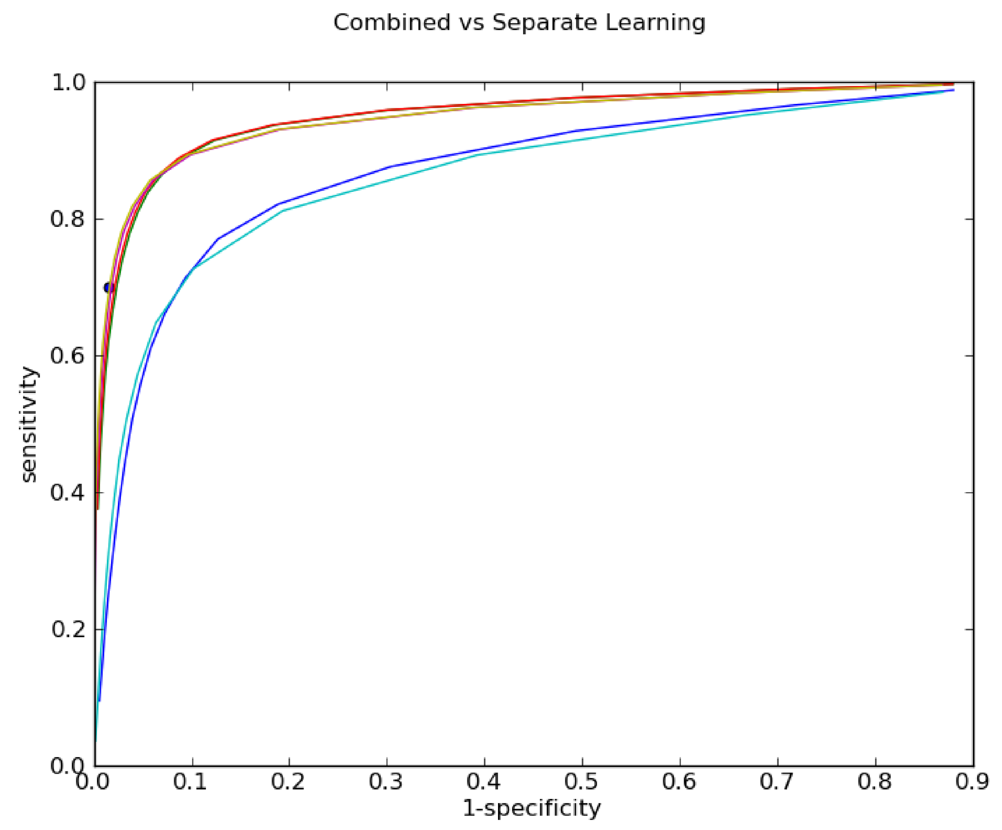


Figure 2: default

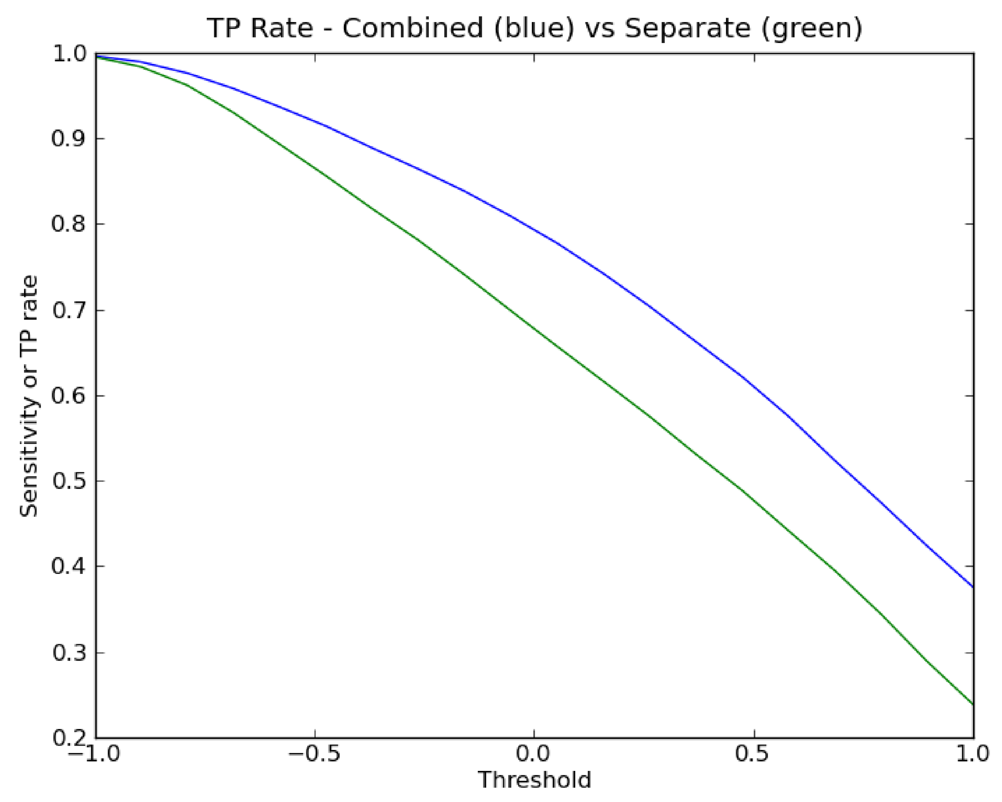


Figure 3: default

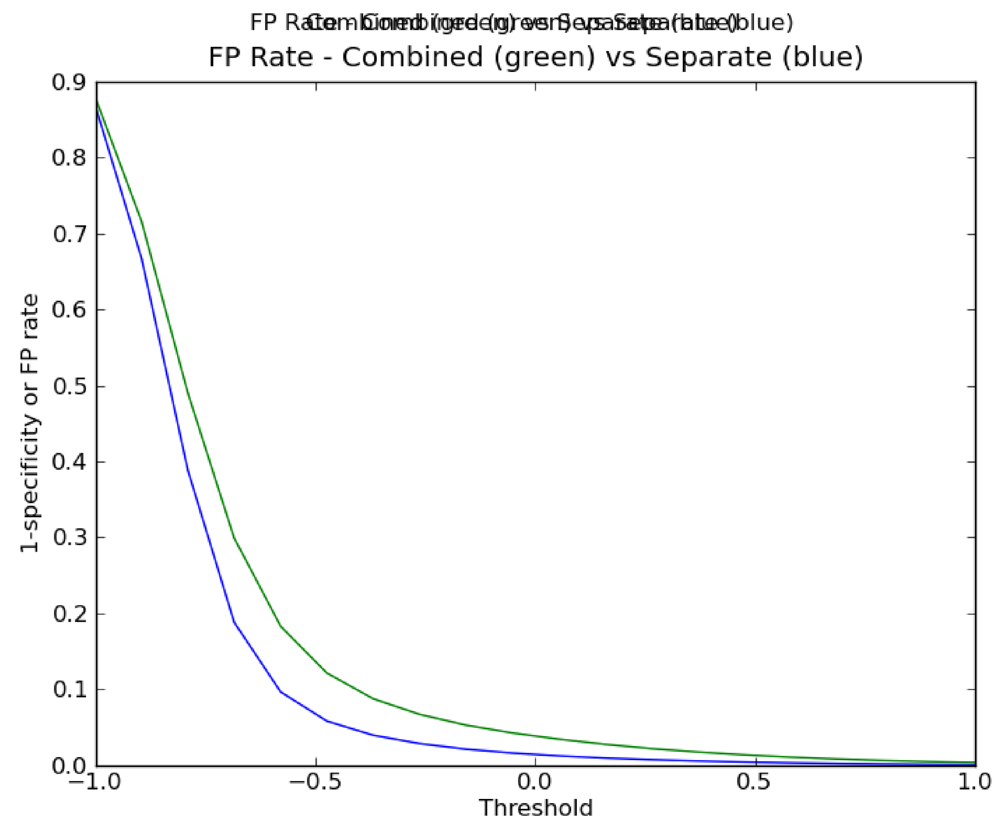


Figure 4: default

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^N p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^K p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^I p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^K\left(\prod_{n=1}^N p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^K p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^I p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^K p\left(\phi_k; \beta\right)\prod_{m=1}^M\left\{p\left(\theta_m; \alpha\right)\prod_{n=1}^N\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^I p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^I p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^I\left[p\left(\eta_i; \mu\right)\prod_{m=1}^M p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^K p\left(\phi_k; \beta\right)\prod_{m=1}^M\prod_{n=1}^N p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_\theta\prod_{m=1}^M p\left(\theta_m; \alpha\right)\prod_{n=1}^N p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{15}$$

$$= \prod_{i=1}^I\left[p\left(\eta_i; \mu\right)\prod_{m=1}^M p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^I\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right)\prod_{m=1}^M h\left(y\right)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\bullet),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\bullet),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\bullet)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\bullet)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA) in the same manner as was done in supervised LDA (SLDA) prior art. We find that the additional supervision signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in medical document labeling and product categorization tasks. Additionally, held-out likelihood

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [?], product descriptions and catalogs (e.g. [?] as available from [?]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of unstructured information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [3] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 2 we introduce hierarchically supervised LDA (HSLDA), in Section 3.4 we review related work, and in Section ?? we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$

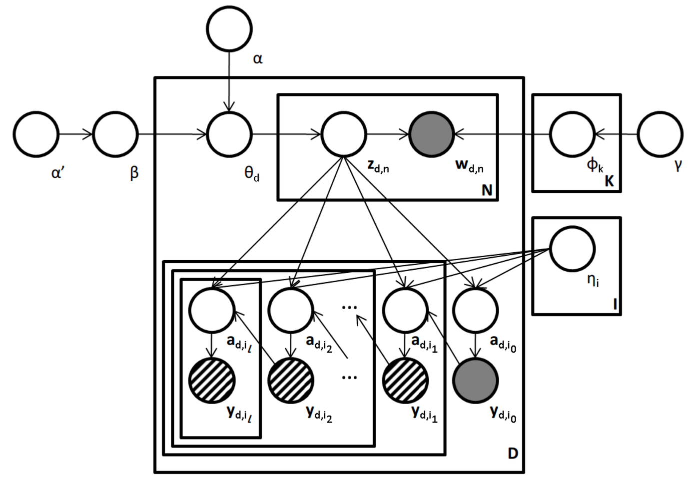


Figure 1: adapted sLDA model

      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases}\mathcal{N}\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1\\ trunc\mathcal{N}^-\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1\end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, \ldots, i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, \ldots, i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases}1, & a_{d,i_{l,c}} > 0\\ -1, & otherwise\end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

2.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

2.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [9]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [9].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

2.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

2.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases}trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1\\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1\end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases}1, & y_{parent_{l,c}} = -1\\ 0.5, & y_{parent_{l,c}} = 1\end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases}\mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1\\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1\\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1\\ 0 & otherwise\end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

2.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [16].

Posterior inference was performed using the "direct assignment" method of Teh et al. [15].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

2.2.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

2.3 Prediction

3 Experiments

3.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

3.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

3.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

5.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

3.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [13, 7, 12, 4], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [11] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [8, 5, 6]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [10]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

3.5 Evaluation

4 Results

5 Discussion

References

[] Amazon, Inc. http://www.amazon.com/, 2011.

[] DMOZ open directory project. http://www.dmoz.org/, 2002.

[1] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[5] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[6] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[7] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[8] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[12] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[13] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[16] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

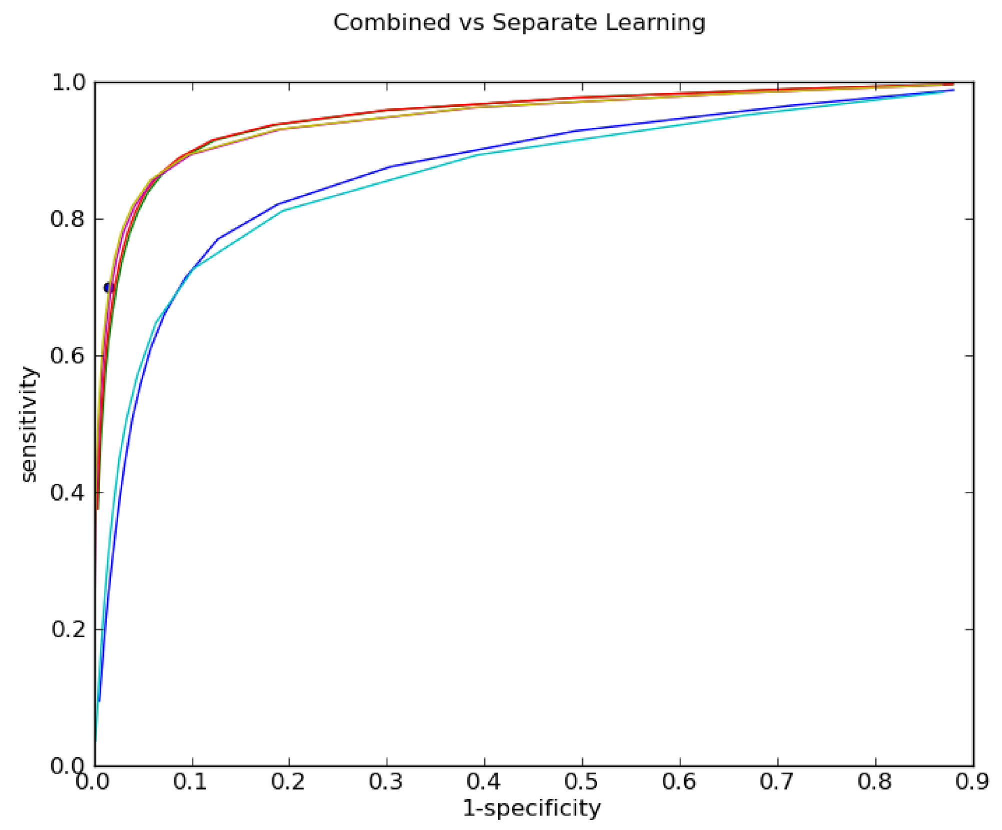


Figure 2: default

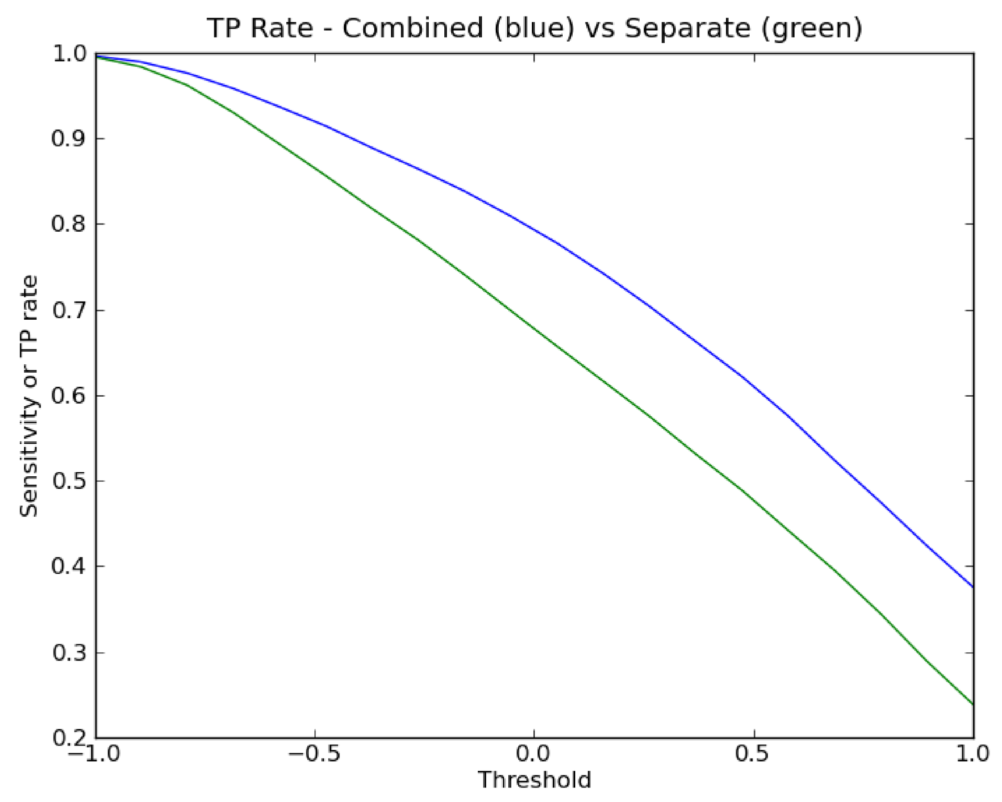


Figure 3: default

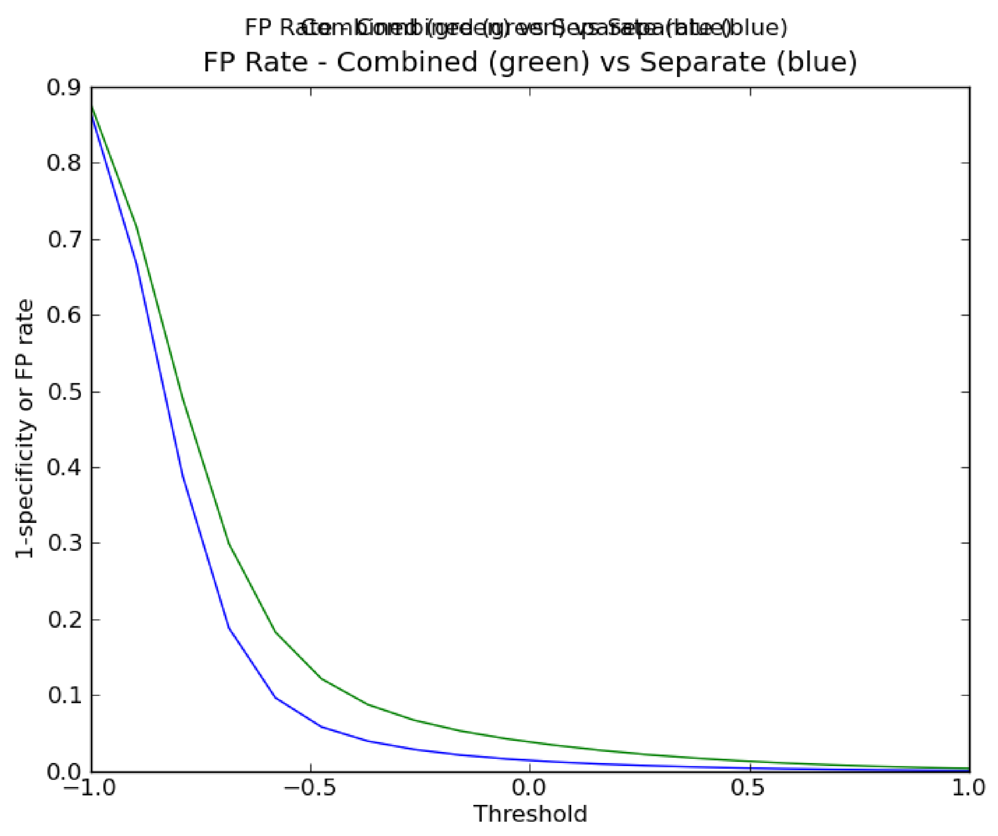


Figure 4: default

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right)\prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right]\prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\}\prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p\left(\phi_k; \beta\right)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p\left(\theta_m; \alpha\right)\prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right)\prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right)\prod_{m=1}^{M} h\left(y\right)\exp\left\{\left(\eta_i^T\bar{z}\right) y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA) in the same manner as was done in supervised LDA (SLDA) prior art. We find that the additional supervision signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in medical document labeling and product categorization tasks. Additionally, held-out likelihood

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 2 we introduce hierarchically supervised LDA (HSLDA), in Section 3.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$

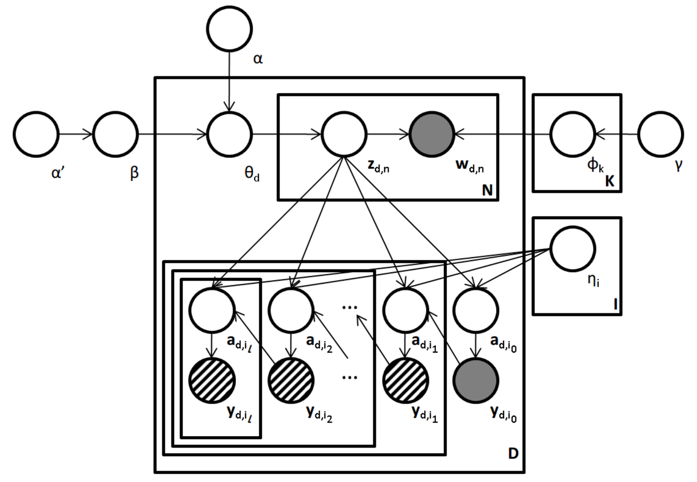


Figure 1: adapted sLDA model

      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^N z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

2.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

2.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^V\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

2.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\boldsymbol{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

2.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right)\prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

2.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

2.2.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

2.3 Prediction

3 Experiments

3.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

3.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

3.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distributed over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

3.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from the distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponentially family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

3.5 Evaluation

4 Results

5 Discussion

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.
[2] DMOZ open directory project. http://www.dmoz.org/, 2002.
[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.
[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.
[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

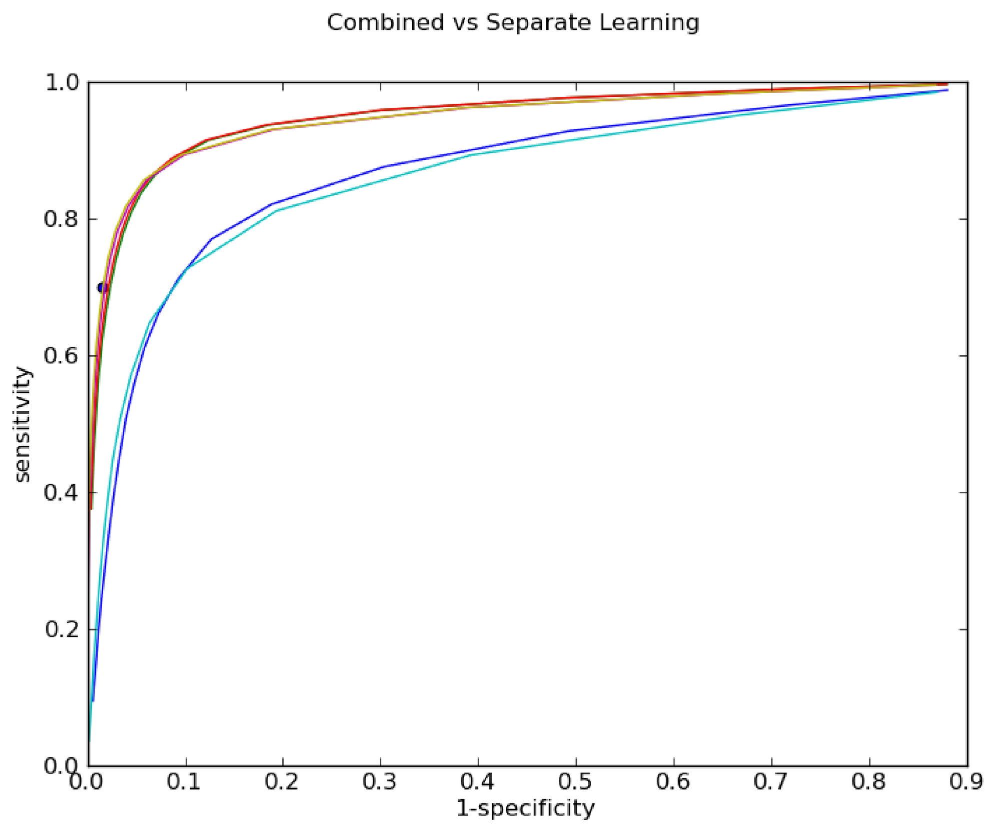


Figure 2: default

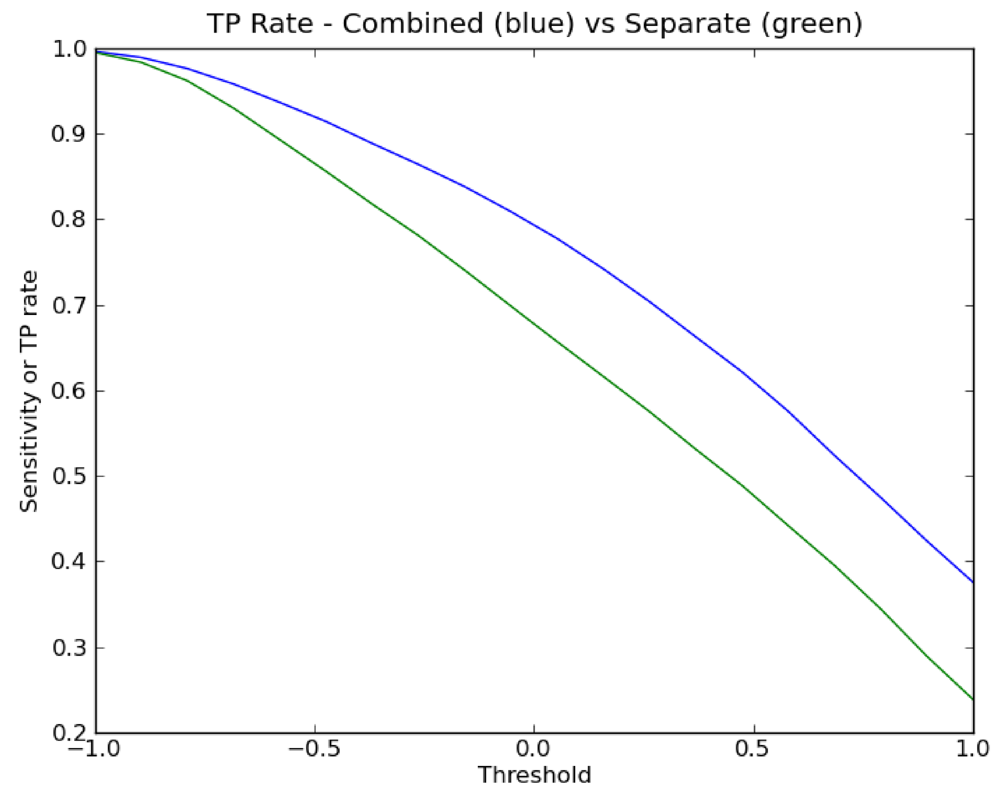


Figure 3: default

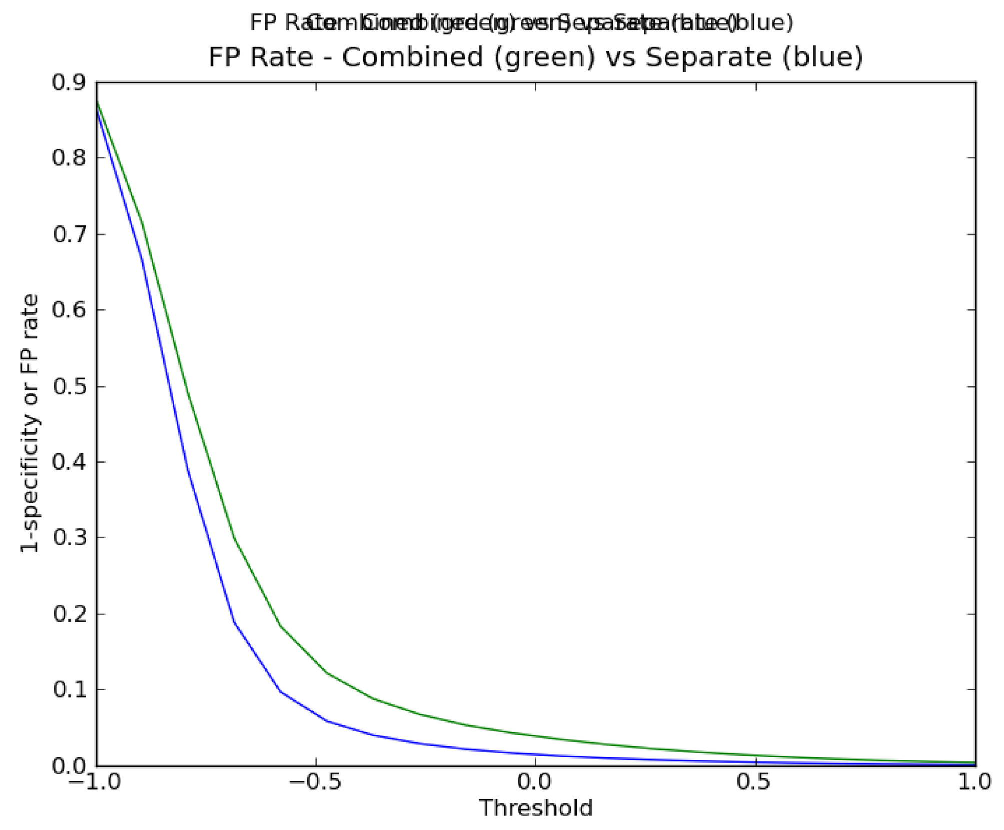


Figure 4: default

5.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^K\left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^K p(\phi_k; \beta)\prod_{m=1}^M\left\{p(\theta_m; \alpha)\prod_{n=1}^N\left[p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \phi_{z_{m,n}})\right]\prod_{i=1}^I p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\}\prod_{i=1}^I p(\eta_i; \mu) d\theta d\phi_{1:K} \tag{14}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^I\left[p(\eta_i; \mu)\prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\int_{\phi_{1:K}}\prod_{k=1}^K p(\phi_k; \beta)\prod_{m=1}^M\prod_{n=1}^N p(w_{m,n} \mid \phi_{z_{m,n}}) d\phi_{1:K}\int_\theta\prod_{m=1}^M p(\theta_m; \alpha)\prod_{n=1}^N p(z_{m,n} \mid \theta_m) d\theta \tag{15}$$

$$= \prod_{i=1}^I\left[p(\eta_i; \mu)\prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma(\beta_v)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma(\alpha_k)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^I\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^M h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V\beta_v\right)}{\prod_{v=1}^V\Gamma(\beta_v)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\bullet),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\bullet),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K\alpha_k\right)}{\prod_{k=1}^K\Gamma(\alpha_k)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\bullet)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\bullet)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA) in the same manner as was done in supervised LDA (SLDA) prior art. We find that the additional supervision signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in medical document labeling and product categorization tasks. Additionally, held-out likelihood

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision," often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 2 we introduce hierarchically supervised LDA (HSLDA), in Section 3.4 we review related work, and in Section ?? we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Methods

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{l_c} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$
   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$

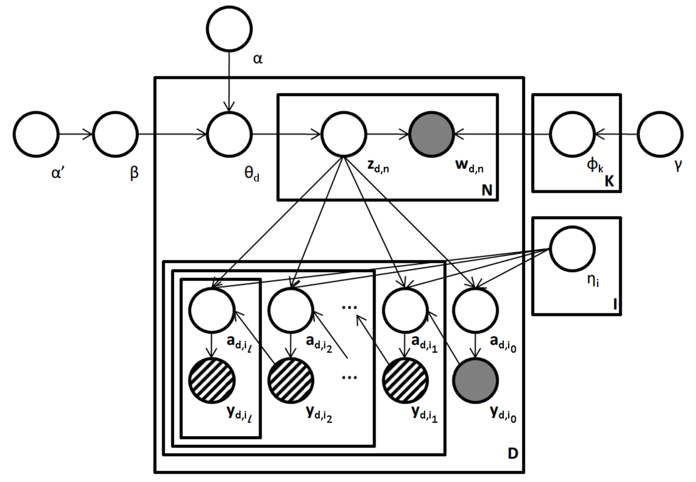


Figure 1: adapted sLDA model

ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
(c) For each level of the ICD-9 code tree, l:
i. For each ICD-9 code at this level, c:
A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^N z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

2.1 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\eta,a,\beta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

2.2 Gibbs Sampling

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.2.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^V\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

2.2.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

2.2.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_{i_{l,c}}^T\bar{z}, 1, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation ??.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,j}\in children_{l,c}} p\left(y_{i_{l,j}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,j}} \in y_{children_{l,c}}, y_{i_{l,j}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,j}} \in y_{children_{l,c}} \setminus y_{i_{l,j}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

2.2.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n, m)$ represents stirling numbers of the first kind.

2.2.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

2.3 Prediction

3 Experiments

3.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

3.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

3.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

3.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

3.5 Evaluation

4 Results

5 Discussion

- what about the nonparametric version of this?

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

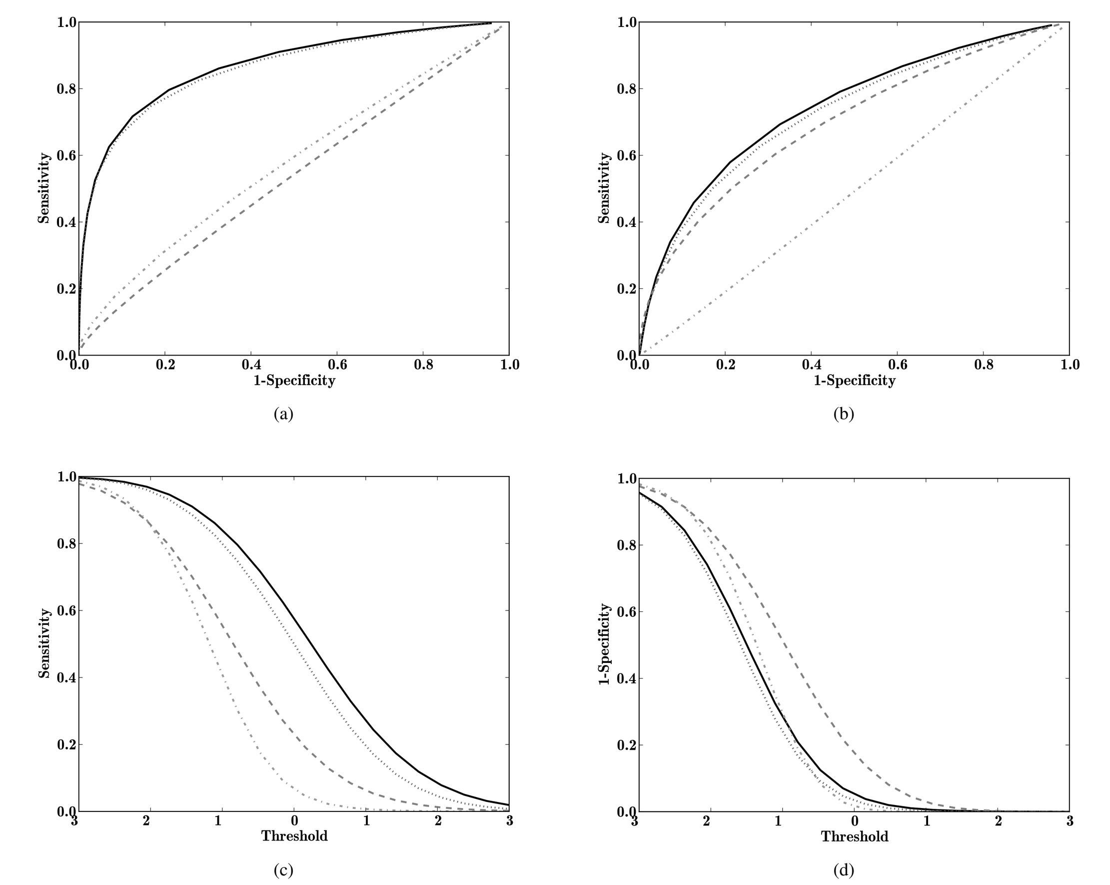


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

5.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^N p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^K\left(\prod_{n=1}^N p(z_n \mid \theta)\, p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^K p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^I p(y_i \mid z_{1:N}, \eta_i, \xi)\, p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi)\, d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^K p(\phi_k; \beta)\prod_{m=1}^M\left\{p(\theta_m; \alpha)\prod_{n=1}^N\left[p(z_{m,n} \mid \theta_m)\, p(w_{m,n} \mid \phi_{z_{m,n}})\right]\prod_{i=1}^I p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\}\prod_{i=1}^I p(\eta_i; \mu)\, d\theta d\phi_{1:K} \tag{14}$$

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi) = \prod_{i=1}^I\left[p(\eta_i; \mu)\prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\int_{\phi_{1:K}}\prod_{k=1}^K p(\phi_k; \beta)\prod_{m=1}^M\prod_{n=1}^N p(w_{m,n} \mid \phi_{z_{m,n}})\, d\phi_{1:K}\int_\theta\prod_{m=1}^M p(\theta_m; \alpha)\prod_{n=1}^N p(z_{m,n} \mid \theta_m)\, d\theta \tag{15}$$

$$= \prod_{i=1}^I\left[p(\eta_i; \mu)\prod_{m=1}^M p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V \beta_v\right)}{\prod_{v=1}^V\Gamma(\beta_v)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K\Gamma(\alpha_k)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^I\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^M h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^K\frac{\Gamma\left(\sum_{v=1}^V \beta_v\right)}{\prod_{v=1}^V\Gamma(\beta_v)}\frac{\prod_{v=1}^V\Gamma\left(n_{(\cdot),v}^k + \beta_v\right)}{\Gamma\left(\sum_{v=1}^V n_{(\cdot),v}^k + \beta_v\right)}\prod_{m=1}^M\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K\Gamma(\alpha_k)}\frac{\prod_{k=1}^K\Gamma\left(n_{m,(\cdot)}^k + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{m,(\cdot)}^k + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 2 we introduce hierarchically supervised LDA (HSLDA), in Section 3.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V\left(\mathbf{1}, \gamma\right)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K\left(-\mathbf{1}, \sigma\right)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K\left(\mathbf{1}, \alpha'\right)$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K\left(\beta, \alpha\right)$

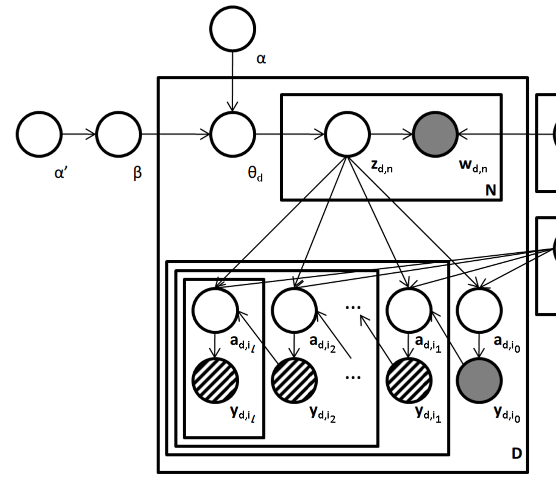


Figure 1: adapted sLDA model

   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K\left(\theta_d\right)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V\left(\beta_{z_n}\right)$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T\eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3 Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.1.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m,n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d,n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d,n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d,n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d,n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.1.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is apporopriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha; \lambda\right)$, $p\left(\alpha'; \lambda\right)$, $p\left(\beta; \lambda\right)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

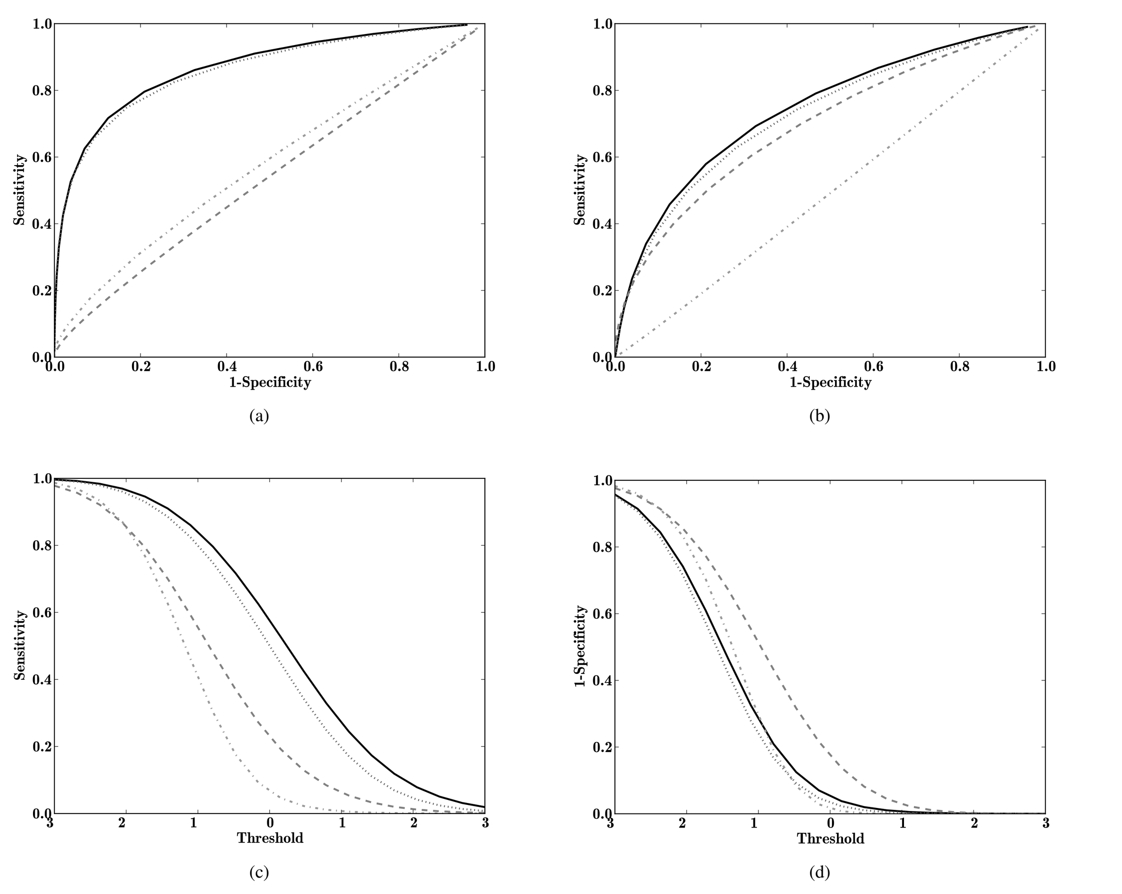


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed is ADLER, dotted is ADLER, and dot-dashed is ADLER.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section **??** we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}_V\gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}\alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta\alpha)$

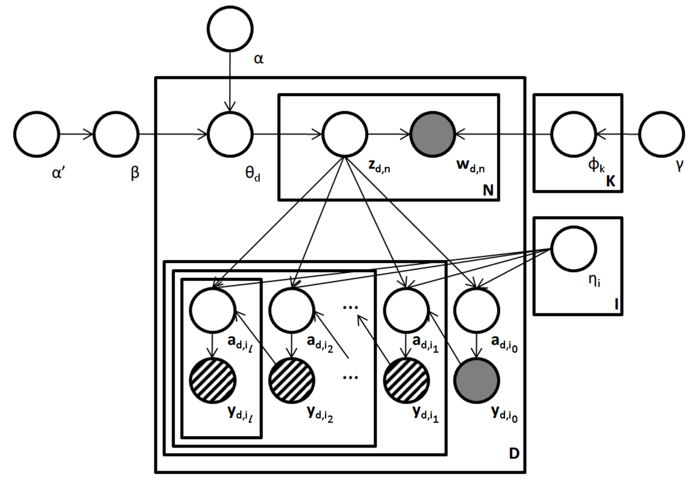


Figure 1: adapted sLDA model

(b) For each word, n:
   i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
   ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_{n}})$
(c) For each level of the ICD-9 code tree, l:
   i. For each ICD-9 code at this level, c:
      A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}(\bar{z}^T\eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-(\bar{z}^T\eta_{i_{l,c}}, 1), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3 Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_{z} p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.1.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.1.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right)\prod_{i_{l,\xi}\in children_{l,c}} p\left(y_{i_{l,\xi}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,\xi}} \in y_{children_{l,c}}, y_{i_{l,\xi}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,\xi}} \in y_{children_{l,c}} \setminus y_{i_{l,\xi}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

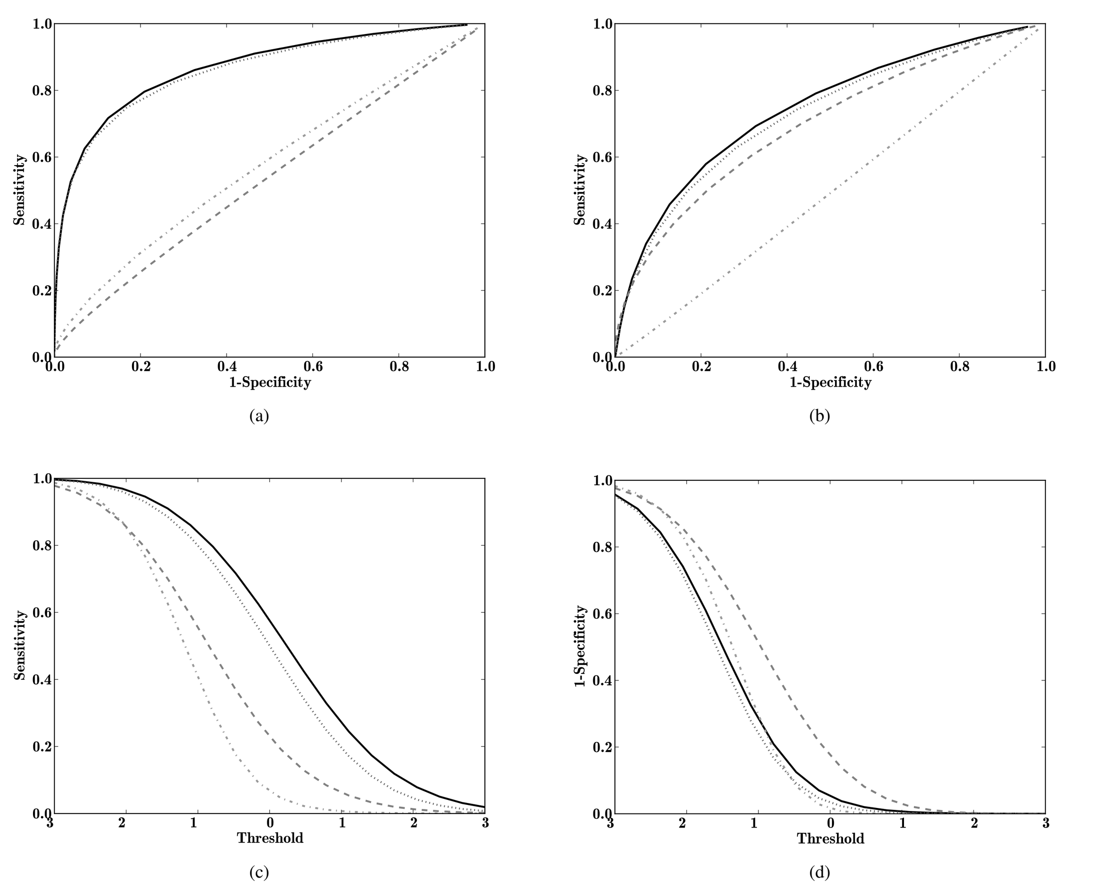


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed is ADLER, dotted is ADLER, and dot-dashed is ADLER.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right)}{\int_\theta p(\theta \mid \alpha)\sum_{k=1}^{K}\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \phi_{1:K})\right)\left(\prod_{k=1}^{K} p(\phi_k \mid \beta)\right)\left(\prod_{i=1}^{I} p(y_i \mid z_{1:N}, \eta_i, \xi) p(\eta_i \mid \mu)\right) d\theta} \tag{13}$$

$$\int_\theta\int_{\phi_{1:K}} p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi) \, d\theta d\phi_{1:K} = \int_\theta\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\left\{p(\theta_m; \alpha)\prod_{n=1}^{N}\left[p(z_{m,n} \mid \theta_m) p(w_{m,n} \mid \phi_{z_{m,n}})\right]\prod_{i=1}^{I} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right\}\prod_{i=1}^{I} p(\eta_i; \mu) \, d\theta d\phi_{1:K} \tag{14}$$

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi) = \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\int_{\phi_{1:K}}\prod_{k=1}^{K} p(\phi_k; \beta)\prod_{m=1}^{M}\prod_{n=1}^{N} p(w_{m,n} \mid \phi_{z_{m,n}}) \, d\phi_{1:K}\int_\theta\prod_{m=1}^{M} p(\theta_m; \alpha)\prod_{n=1}^{N} p(z_{m,n} \mid \theta_m) \, d\theta \tag{15}$$

$$= \prod_{i=1}^{I}\left[p(\eta_i; \mu)\prod_{m=1}^{M} p(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi)\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\cdot),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\cdot),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\cdot)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\cdot)}^{k} + \alpha_k\right)} \tag{16}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}(\eta_i \mid \mu, 1)\prod_{m=1}^{M} h(y)\exp\left\{\left(\eta_i^T\bar{z}\right)y - A\left(\eta_i^T\bar{z}\right)\right\}\right]\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{17}$$

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section **??** we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$

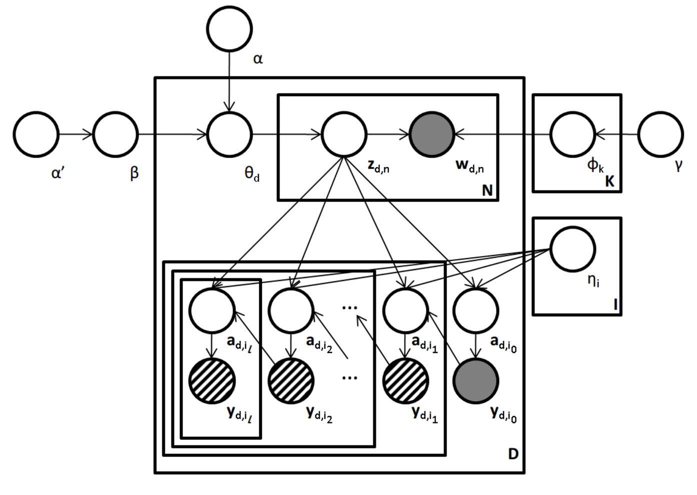


Figure 1: adapted sLDA model

   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $i_l = \{i_{l,0}, i_{l,1}, ..., i_{l,C_l}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3 Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,a,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

3.1.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\mathbf{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.1.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta_i^T\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parents_{l,c}}\right) \prod_{i_{\hat{l},\hat{c}}\in children_{l,c}} p\left(y_{i_{\hat{l},\hat{c}}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}}, y_{i_{\hat{l},\hat{c}}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{\hat{l},\hat{c}}} \in y_{children_{l,c}} \setminus y_{i_{\hat{l},\hat{c}}} = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha; \lambda)$, $p(\alpha'; \lambda)$, $p(\beta; \lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical information, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.
[2] DMOZ open directory project. http://www.dmoz.org/, 2002.
[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.
[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.
[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.
[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.
[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.
[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.
[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.
[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.
[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.
[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

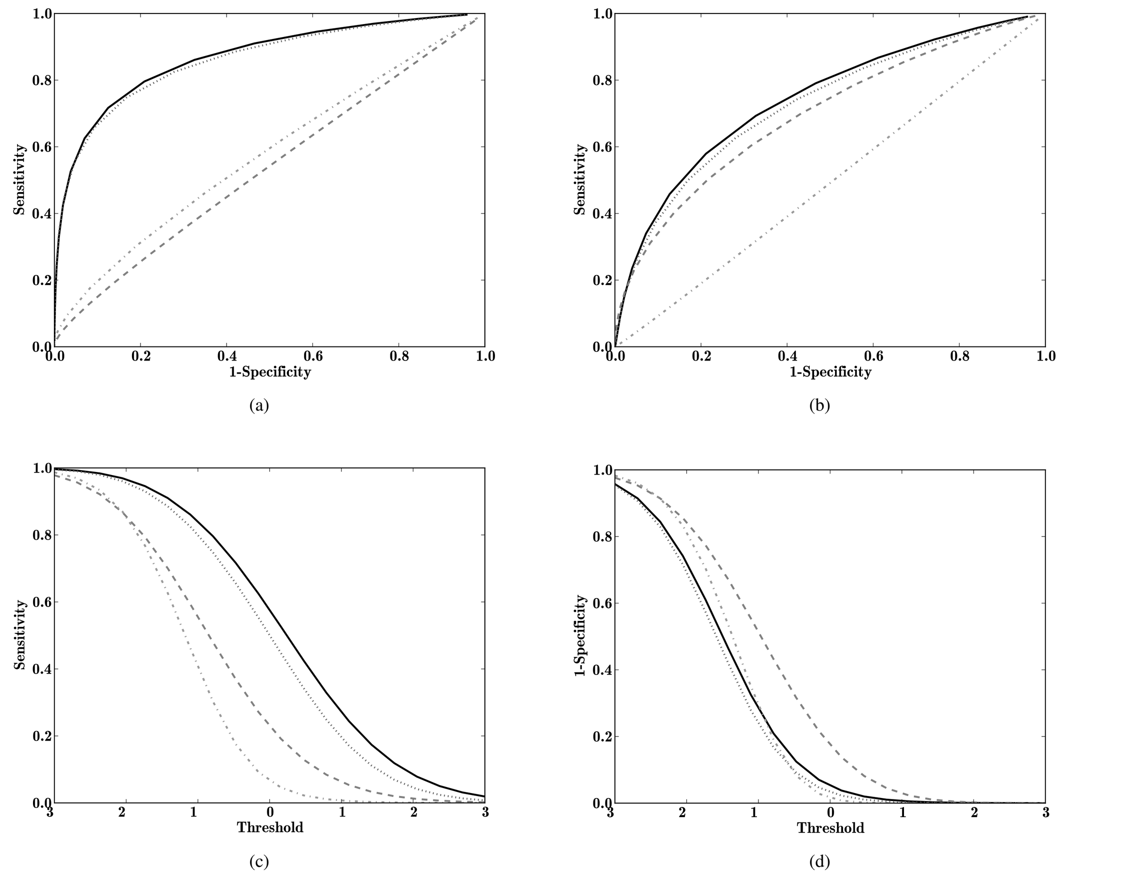


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: **??** includes ancestor prediction performance, **??** results are for given (leaf) labels alone. Bottom row: **??** are the sensitivity curves from **??** aligned on threshold value, **??** are the 1-specificity curves from **??** aligned on threshold value.

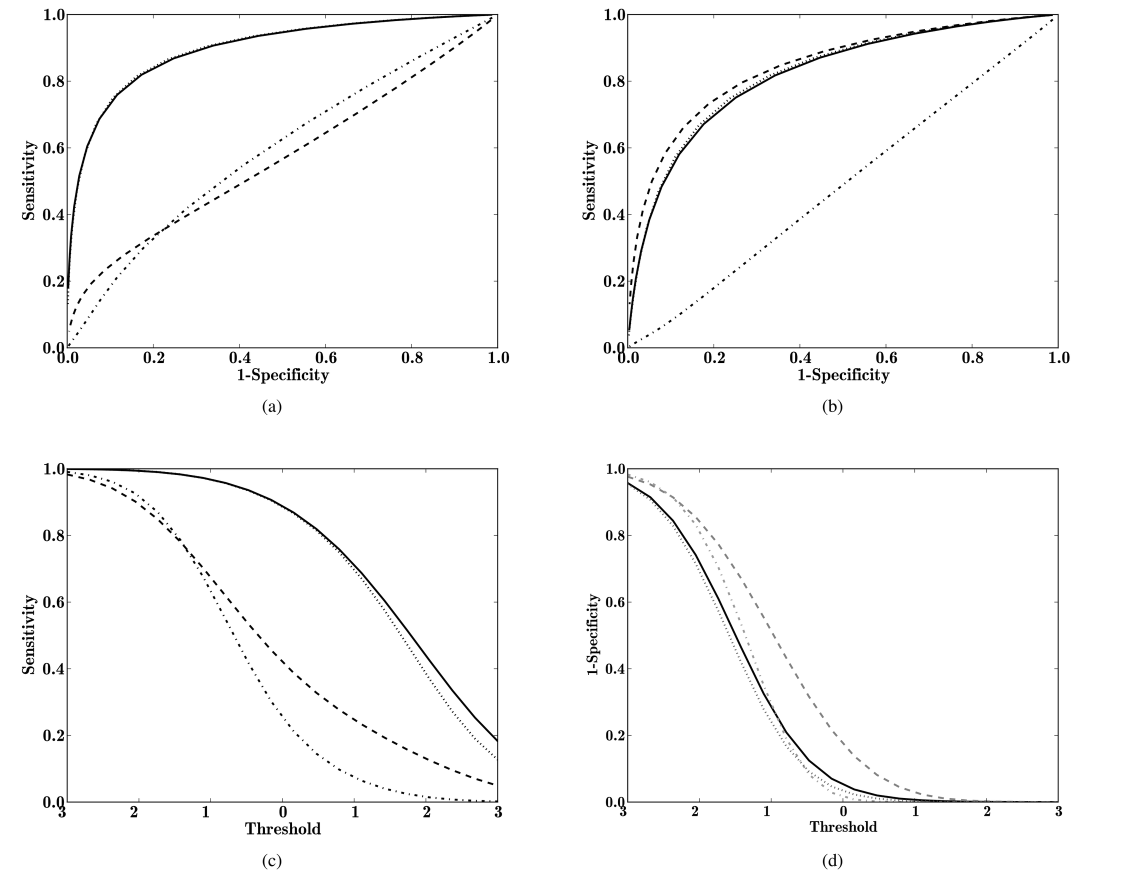


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: **??** includes ancestor prediction performance, **??** results are for given (leaf) labels alone. Bottom row: **??** are the sensitivity curves from **??** aligned on threshold value, **??** are the 1-specificity curves from **??** aligned on threshold value.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.
[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section **??** we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. We assume a pre-specified set of labels $\mathcal{L}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many, label(s) in $\mathcal{L}$. A response of 1 or -1 to label l indicates if the document respectively is or is not l. The label l for a document d will be used interchangeably to refer the observed response of document d to label l. The label set is assumed to be an "is a" hieararchy. This means that if a label l_1 is a parent of label l_2 in the hierarchy and document d has a positive response to the label l_2 then document d also has a positive response to the label l_1. Seen in a negative light, if document d has a negative response to the l_1 label then document d also has a negative response to the l_2 label. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models. Approximate inference is performed using Gibbs sampling.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, and the standard deviation σ used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_k(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

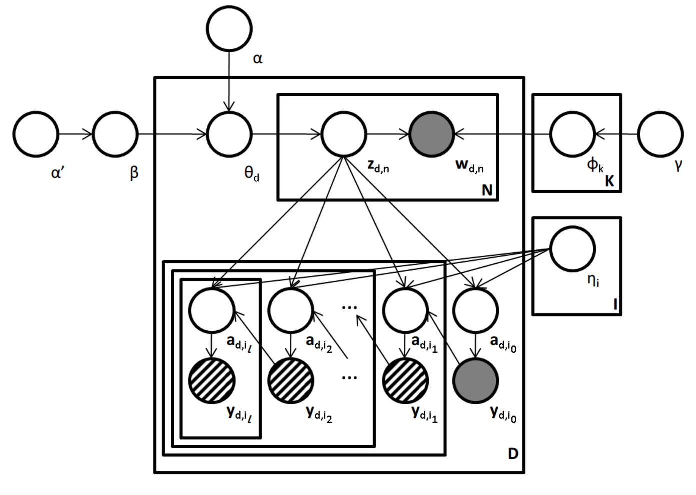


Figure 1: adapted sLDA model

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$:
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim \mathrm{Multinomial}(\beta_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l \sim \mathcal{N}(\bar{z}_d^T \eta_l, 1)$, where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N_d} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{parent(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression is like logistic regression except instead of modeling the logit of $P(y = 1)$ using a linear form we model $\Phi^{-1}(P(y = 1))$ using a linear form, where $\Phi(\cdot)$ is the CDF for a standard normal distribution. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

3 Inference

In the Bayesian approach to statistical modeling the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ as in Griffiths and Steyvers [12]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\eta = \{\eta_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l,d}\}_{l \in \mathcal{L}, d=1,\ldots,D}$, β, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \backslash z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [12]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket, explicitly this means

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant [12] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \quad (2)$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 4, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients **etal** for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3)$$

$$\hat{\mu} = \hat{\Sigma}_l \left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(a_{l,d} - {\eta_l}^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics [18].

Posterior inference is performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \quad (6)$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \quad (7)$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha)$, $p(\alpha')$, $p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(2, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

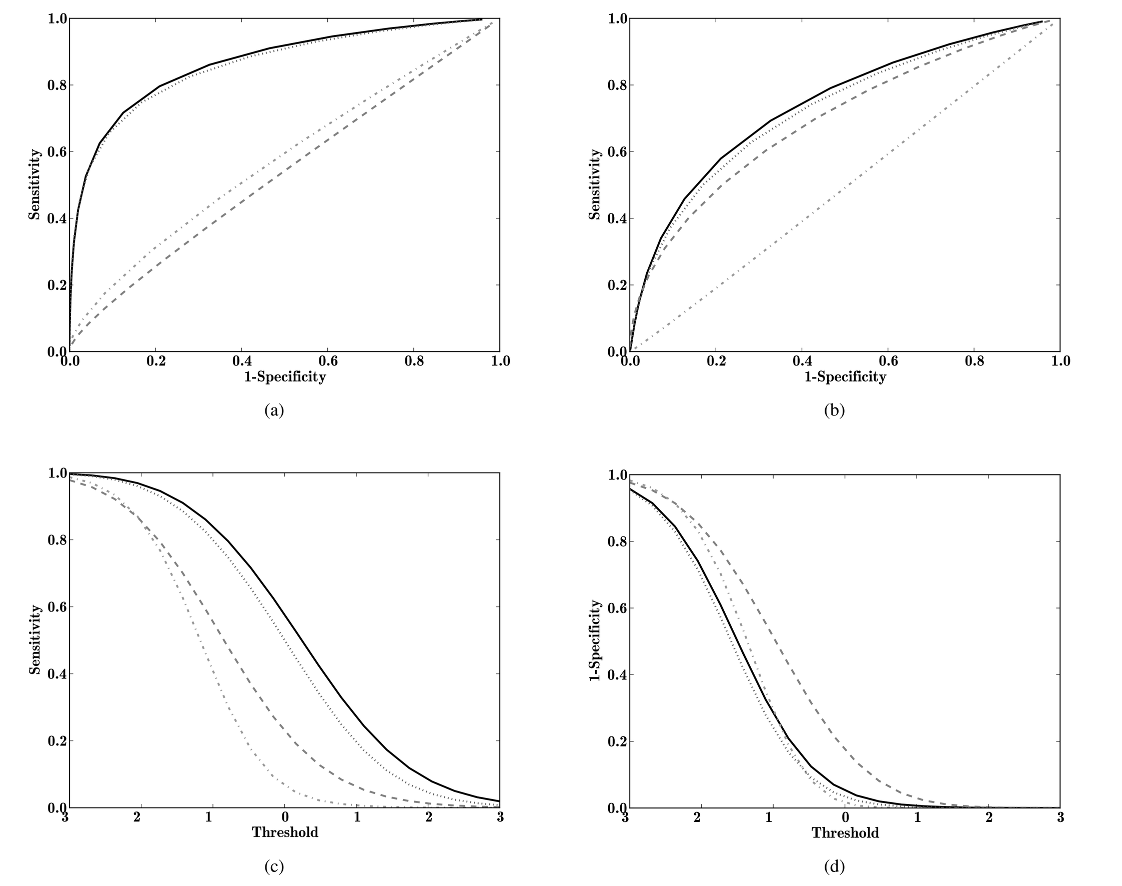


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

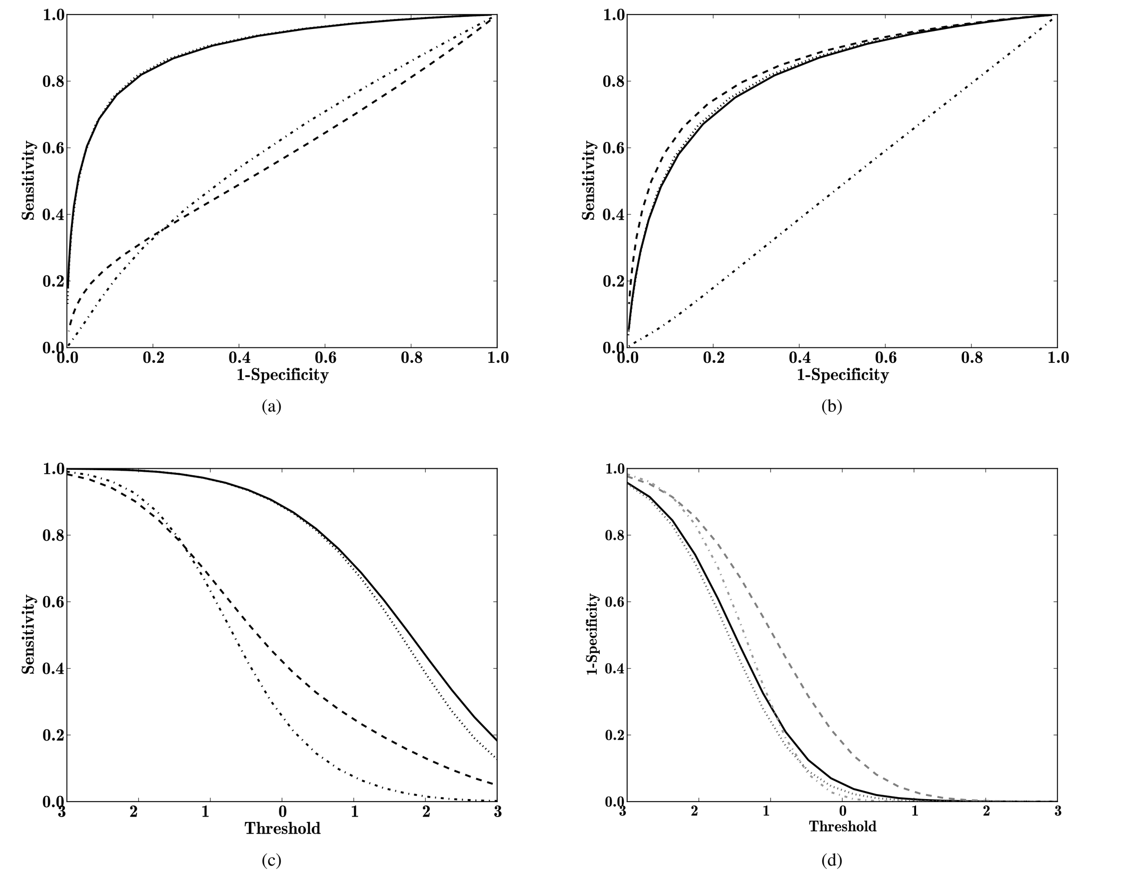


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section **??** we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

Given the number of topics, K, and broad gamma priors on hyperparameters, the generative process is as follows:

1. For each topic, k:
   (a) Draw a distribution over words $\phi_k \sim Dir_V(\mathbf{1}, \gamma)$
2. For each ICD9 code, c, at all levels in the tree, l:
   (a) Draw regression coefficient $\eta_{i_{l,c}} \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma)$
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim Dir_K(\mathbf{1}, \alpha')$
4. For each document, d:
   (a) Draw topic proportions $\theta_d \mid \beta, \alpha \sim Dir_K(\beta, \alpha)$

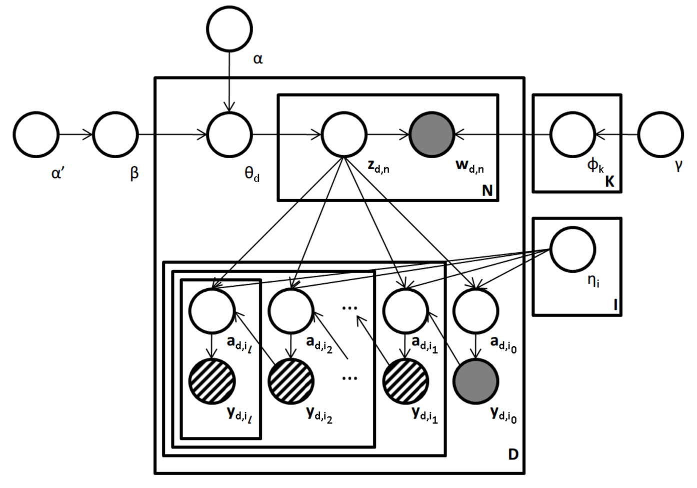


Figure 1: adapted sLDA model

   (b) For each word, n:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult_K(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult_V(\beta_{z_n})$
   (c) For each level of the ICD-9 code tree, l:
      i. For each ICD-9 code at this level, c:
         A. Draw a latent variable

$$a_{d,i_{l,c}} \sim \begin{cases} \mathcal{N}\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(\bar{z}^T \eta_{i_{l,c}}, 1\right), & y_{parent_{l,c}} = -1 \end{cases}$$

where $\bar{z} = N^{-1}\sum_{n=1}^{N} z_n$ and $\mathcal{I} = \{i_0, i_1, ..., i_{\mathcal{L}}\}$ and $\mathcal{I} = \{i_{l,0}, i_{l,1}, ..., i_{l,|\mathcal{C}|}\}$

         B. Draw a response variable $y_{d,i_{l,c}} \mid a_{d,i_{l,c}} \sim \begin{cases} 1, & a_{d,i_{l,c}} > 0 \\ -1, & otherwise \end{cases}$

The generative model for the ICD-9 codes is equivalent to a probit regression model. In our case, the regression is conditional on the parents according to the constraints of the ICD-9 code tree. The latent variable utilized here is also known as an auxiliary variable.

3 Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation .

$$p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma \mid w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right) \tag{1}$$

$$= \frac{p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}{\int_{\theta,\phi,\alpha,\eta,\alpha,\alpha',\beta,\gamma}\sum_z p\left(\theta, z_{1:N}, \phi_{1:K}, \eta_{i_{l,c}\in\mathcal{I}}, a_{i_{l,c}\in\mathcal{I}}, \beta, \alpha, \alpha', \gamma, w_{1:N}, y_{i_{l,c}\in\mathcal{I}}; \sigma, \lambda\right)}$$

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model, marginalizing $\theta_{1:D}$ and $\phi_{1:K}$. For details regarding collapsing in LDA models see Griffiths and Steyvers [12]. To derive the Gibbs sampler we have to evaluate the individual conditional probability distributions for all unobserved variables.

3.1.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(m,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, d. The conditional probability with respect to this latent variable is proportional to the joint distribution of its markov blanket up to a constant.

$$p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto p\left(z_{d,n}, \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \tag{2}$$

Due to the factorization of this model, we can rewrite the joint distribution as the following:

$$\propto \prod_{i_{l,c}\in\mathcal{I}} p\left(a_{i_{l,c}} \mid \mathbf{z}, \eta_{i_{l,c}}\right) p\left(z_{d,n}, \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right) \tag{3}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [12].

$$\propto \prod_{i_{l,c}\in\mathcal{I}} \exp\left\{-\frac{\left(\bar{z}^T\eta_{i_{l,c}} - a_{i_{l,c}}\right)^2}{2}\right\}\left(n_{d,(\cdot)}^{k,-(d,n)} + \alpha\beta_k\right)\frac{n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma}{\sum_{v=1}^{V}\left(n_{(\cdot),w_{d,n}}^{k,-(d,n)} + \gamma\right)} \tag{4}$$

Here, $n_{d,v}^{k,-(d,n)}$ represents the count of word v in document d assigned to topic k omitting the $(d,n)^{th}$ word count.

Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right)$

Given that $\eta_{i_{l,c}}$ and $a_{d,i_{l,c}}$ are distributed normally, this posterior distribution is also normal. We evaluated the model over various values of σ where $\sigma = \{0.01, 0.1, 0.25, 1, 2\}$.

$$p\left(\eta_{i_{l,c}} \mid \mathbf{z}_{1:D}, \mathbf{a}; \sigma\right) = \mathcal{N}\left(\eta_{i_{l,c}} \mid \hat{\mu}_i, \hat{\Sigma}_i\right) \tag{5}$$

$$\hat{\mu}_i = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_{(\cdot),i_{l,c}}\right)$$

$$\hat{\Sigma}_i^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}$$

3.1.3 $p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$ and $p\left(y_{m,i} \mid \mathbf{a}\right)$

In the augmented probit regression model, the posterior distribution of $a_{i_{l,c}}$ is distributed according to a truncated normal distribution where the response variable is observed.

$$p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) = \begin{cases} trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \eta^T_{i_{l,c}}\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = 1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \eta^T_{i_{l,c}}\bar{z}, \mathbf{1}, y_{d,i_{l,c}}\right) & if \quad y_{d,i_{l,c}} = -1 \end{cases} \tag{6}$$

However, if $y_{d,i_{l,c}}$ is unobserved then $a_{d,i_{l,c}}$ must be sampled jointly with $y_{d,i_{l,c}}$ to ensure that the Markov chain is ergodic. Suppose that $a_{d,i_{l,c}}$ is sampled to have a negative value and $y_{d,i_{l,c}}$ is appropriately sampled at -1. Although there exist valid states where $a_{d,i_{l,c}} > 0$ and $y_{d,i_{l,c}} = 1$, they will never be reached by such a Markov chain since $p\left(a_{d,i_{l,c}} < 0 \mid y_{d,i_{l,c}} = -1\right) = 1$ and $p\left(y_{d,i_{l,c}} = -1 \mid a_{d,i_{l,c}} < 0\right) = 1$. Therefore, to ensure ergodicity, $a_{d,i_{l,c}}$ and $y_{d,i_{l,c}}$ must be sampled from the joint distribution as shown in Equation **??**.

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right) \propto p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) p\left(a_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \tag{7}$$

$$p\left(y_{i_{l,c}} \mid \mathbf{a}, \mathbf{y}_{-(l,c)}\right) = \delta\left(sign\left(a_{d,i_{l,c}}\right) = y_{i_{l,c}}\right) p\left(y_{i_{l,c}} \mid y_{parent_{l,c}}\right) \prod_{i_{l,c}\in children_{l,c}} p\left(y_{i_{l,c}} \mid y_{i_{l,c}}\right) \tag{8}$$

$$p\left(y_{i_{l,c}} = -1 \mid y_{parent_{l,c}}\right) = \begin{cases} 1, & y_{parent_{l,c}} = -1 \\ 0.5, & y_{parent_{l,c}} = 1 \end{cases} \tag{9}$$

$$p\left(a_{d,i_{l,c}}, y_{d,i_{l,c}} \mid \mathbf{z}, \mathbf{Y}_{-(d,i_{l,c})}, \eta\right)$$

$$= \begin{cases} \mathcal{N}\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right) p\left(y_{d,i_{l,c}} \mid a_{d,i_{l,c}}\right), & y_{parent_{l,c}} = 1, \forall y_{i_{l,c}} \in y_{children_{l,c}}, y_{i_{l,c}} = -1 \\ trunc\mathcal{N}^-\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = -1\right), & y_{parent_{l,c}} = -1 \\ trunc\mathcal{N}^+\left(a_{d,i_{l,c}} \mid \bar{z}^T\eta_{i_{l,c}}, 1\right)\delta\left(y_{d,i_{l,c}} = 1\right), & \exists y_{i_{l,c}} \in y_{children_{l,c}} \setminus y_{i_{l,c}} = 1 \\ 0, & otherwise \end{cases} \tag{10}$$

where $\mathbf{Y}_{-(d,i_{l,c})}$ denotes all of the response variables excluding the response variable being sampled.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics Wallach et al. [18].

Posterior inference was performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim Dir\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{11}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{12}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha;\lambda)$, $p(\alpha';\lambda)$, $p(\beta;\lambda)$

All hyperparameters were given broad gamma priors ($\lambda = \{shape = 2, scale = 1000\}$) and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors

(ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

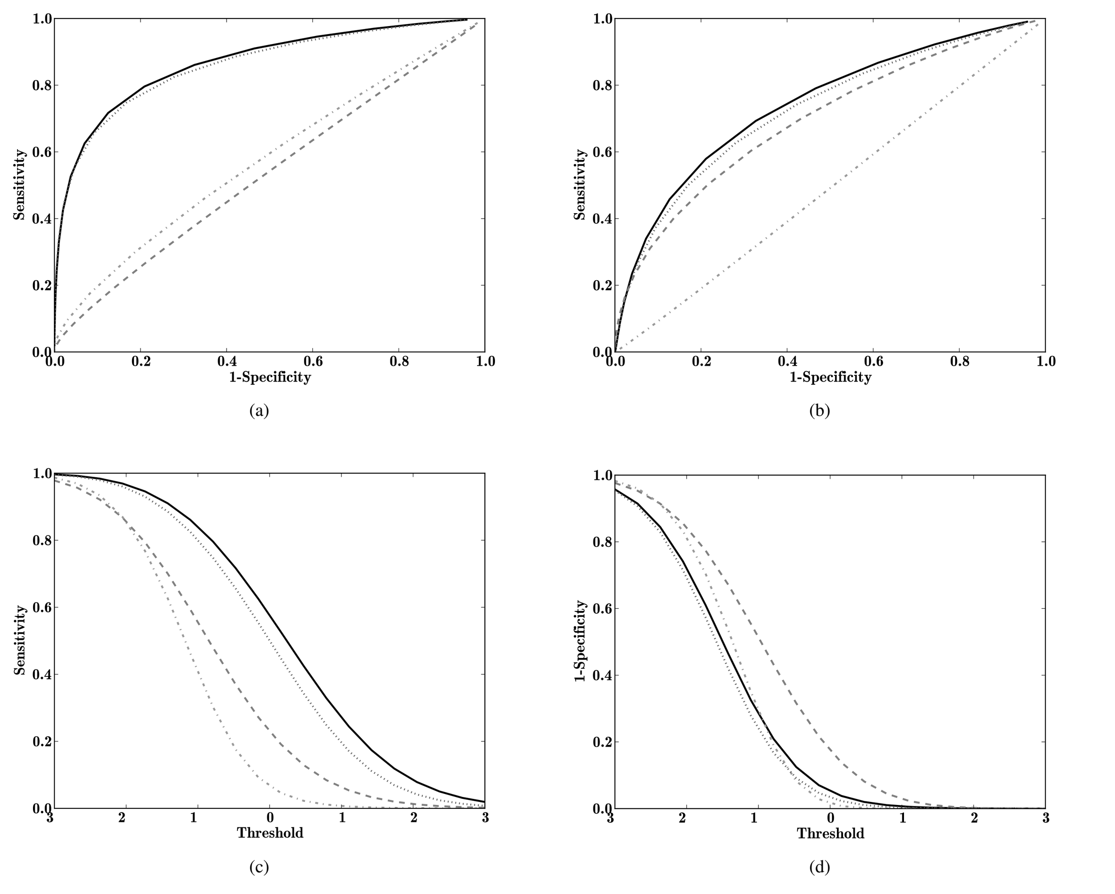


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

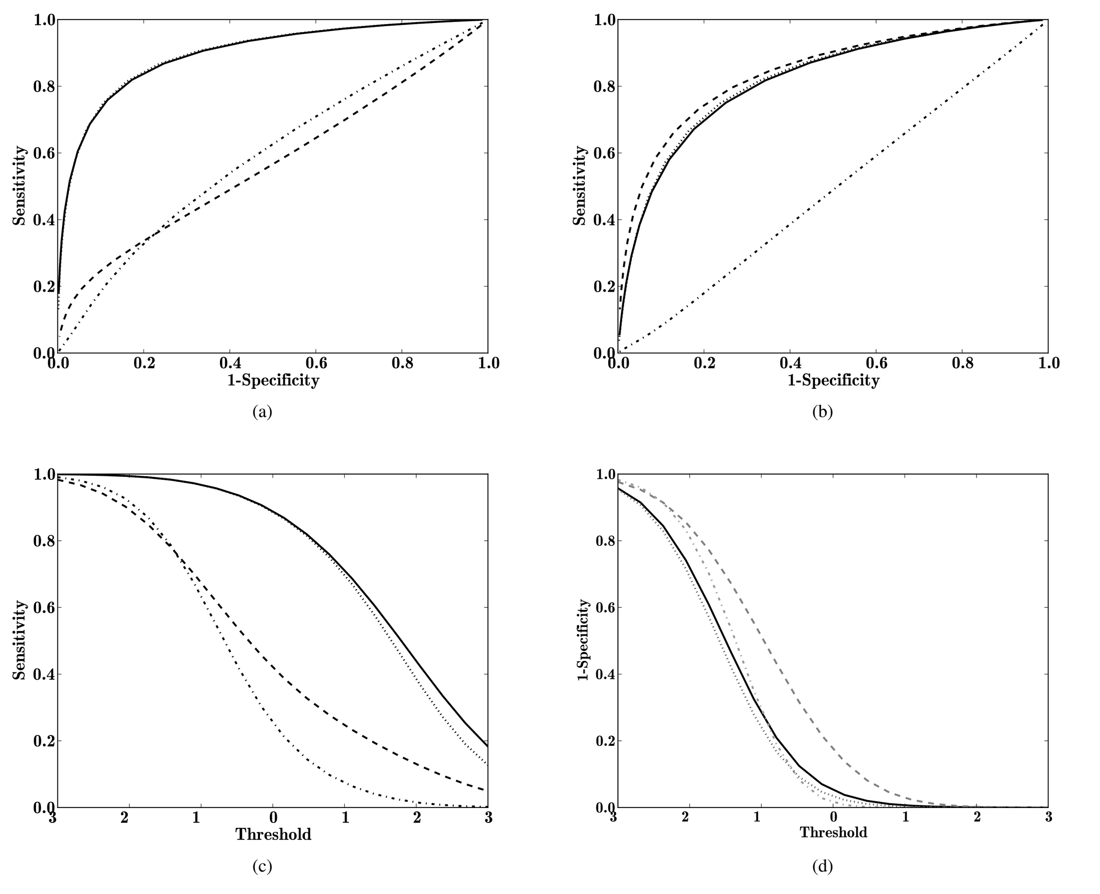


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observe big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section ?? we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. We assume a pre-specified set of labels $\mathcal{L}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many, label(s) in $\mathcal{L}$. A response of 1 or -1 to label l indicates if the document respectively is or is not l. The label l for a document d will be used interchangeably to refer the observed response of document d to label l. The label set is assumed to be an "is a" hierarchy. This means that if a label l_1 is a parent of label l_2 in the hierarchy and document d has a positive response to the label l_2 then document d also has a positive response to the label l_1. Seen in a negative light, if document d has a negative response to the l_1 label then document d also has a negative response to the l_2 label. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models. Approximate inference is performed using Gibbs sampling.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, and the standard deviation σ in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\text{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

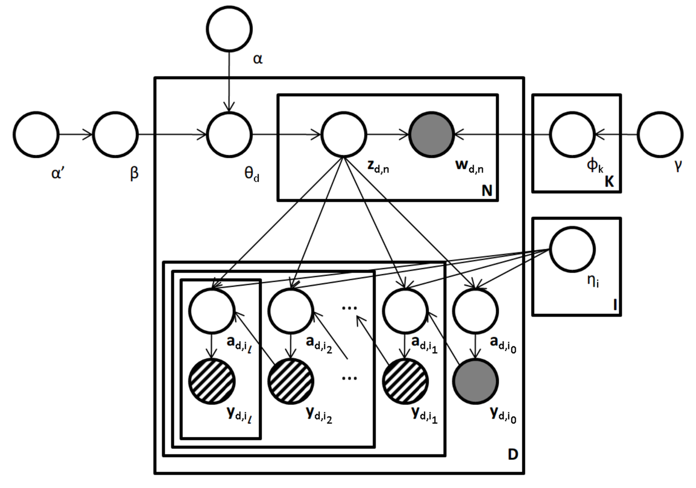


Figure 1: adapted sLDA model

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \text{Dir}_V(\gamma\mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \text{Dir}_K(\alpha'\mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \beta, \alpha \sim \text{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$:
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim \text{Multinomial}(\beta_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l \sim \mathcal{N}(\bar{z}_d^T \eta_l, 1)$, where $\bar{z}_d = N_d^{-1}\sum_{n=1}^{N_d} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{parent(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression is like logistic regression except instead of modeling the logit of $P(y=1)$ using a linear form we model $\Phi^{-1}(P(y=1))$ using a linear form, where $\Phi(\cdot)$ is the CDF for a standard normal distribution. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

3 Inference

In the Bayesian approach to statistical modeling the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [12]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\eta = \{\eta_l\}_{l\in\mathcal{L}}$, $\mathbf{a} = \{a_{l,d}\}_{l\in\mathcal{L},d=1,\ldots,D}$, β, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [12]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket, explicitly this means

$$p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \prod_{l\in\mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \eta_l)\, p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant [12] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l\in\mathcal{L}_d} \exp\left\{-\frac{(\bar{z}_d^T\eta_l - a_{l,d})^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 4, $p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\mathbf{eta_l}$ for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(a_{l,d} - \eta_{\mathbf{l}}{}^T\bar{\mathbf{z}}_d\right)\right\} I(a_{l,d}y_{l,d} > 0). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\beta \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics [18].

Posterior inference is performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim \text{Dir}\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m)(\alpha\beta_k)^m \tag{7}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\text{Gamma}(2, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

4.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for

epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

4.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patient's visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

4.3 ICD-9 Code Hierarchy

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [3]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [3]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

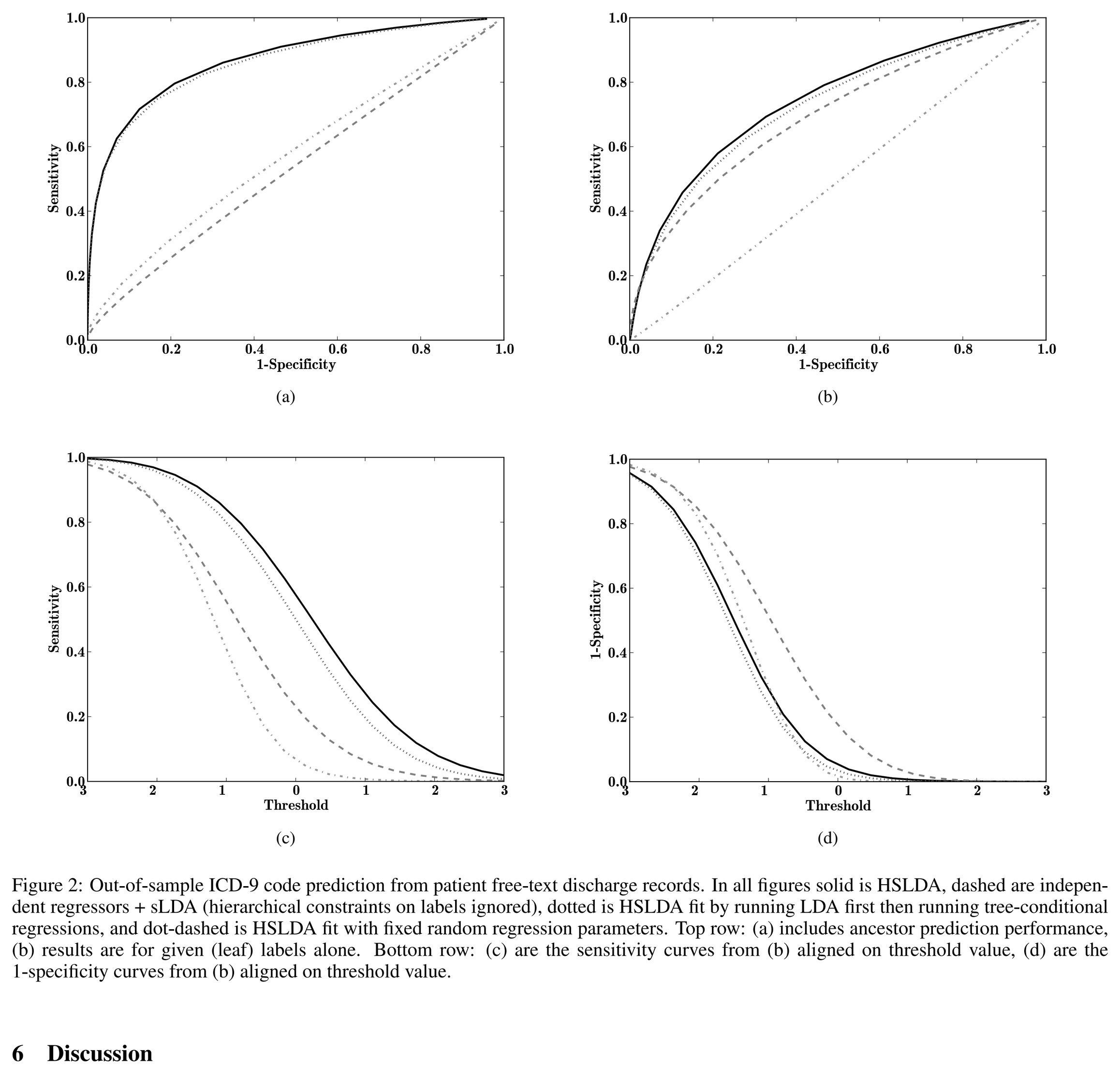


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

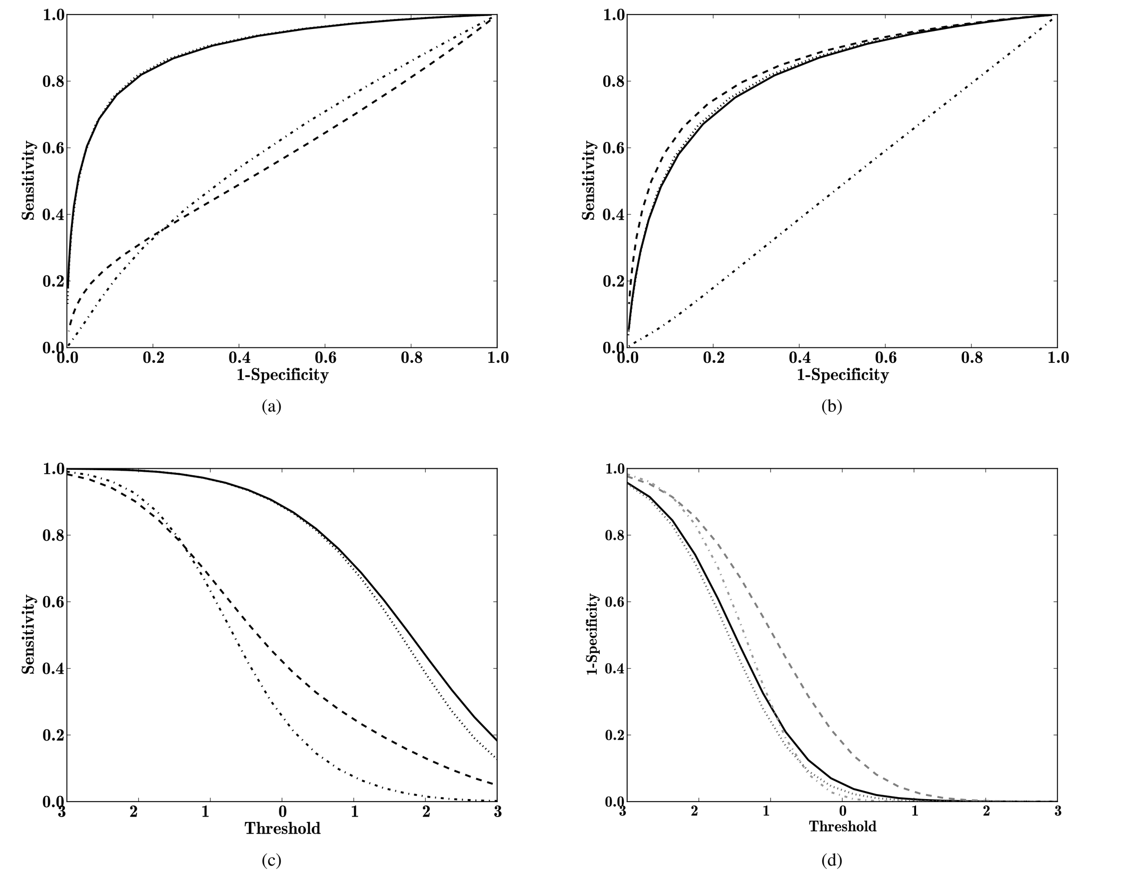


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

(8)

(9)

(10)

(11)

(12)

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section **??** we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. We assume a pre-specified set of labels $\mathcal{L}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many, label(s) in $\mathcal{L}$. A response of 1 or -1 to label l indicates if the document respectively is or is not l. The label l for a document d will be used interchangeably to refer the observed response of document d to label l. The label set is assumed to be an "is a" hieararchy. This means that if a label l_1 is a parent of label l_2 in the hierarchy and document d has a positive response to the label l_2 then document d also has a positive response to the label l_1. Seen in a negative light, if document d has a negative response to the l_1 label then document d also has a negative response to the l_2 label. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models. Approximate inference is performed using Gibbs sampling.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, and the standard deviation σ used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_k(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

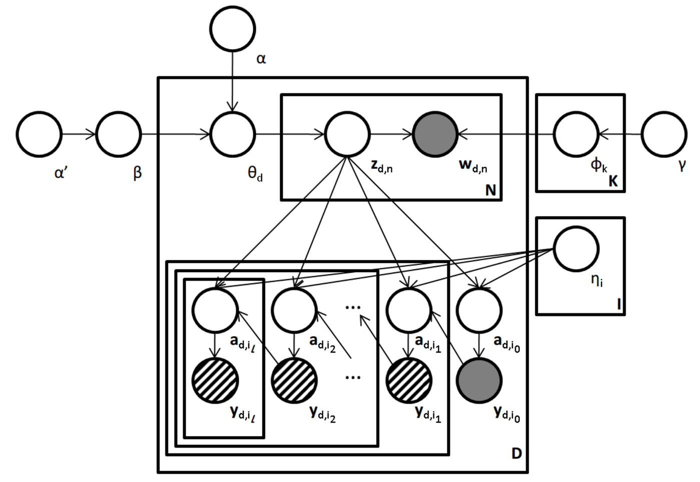


Figure 1: adapted sLDA model

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$:
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l \sim \mathcal{N}(\bar{z}_d^T \eta_l, 1)$, where $\bar{z}_d = N_d^{-1}\sum_{n=1}^{N_d} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{parent(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression is like logistic regression except instead of modeling the logit of $P(y = 1)$ using a linear form we model $\Phi^{-1}(P(y = 1))$ using a linear form, where $\Phi(\cdot)$ is the CDF for a standard normal distribution. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

3 Inference

In the Bayesian approach to statistical modeling the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for every document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L}\setminus\mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [12]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\eta = \{\eta_l\}_{l\in\mathcal{L}}$, $\mathbf{a} = \{a_{l,d}\}_{l\in\mathcal{L},d=1,\ldots,D}$, β, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}\setminus z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [12]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket, explicitly this means

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l\in\mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant [12] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{n^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(n^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l\in\mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T\eta_l - a_{l,d}\right)^2}{2}\right\}. \quad (2)$$

Here, $n^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_{\mathbf{l}} \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients $\mathbf{eta_l}$ for $l \in \mathcal{L}$. Given that $\eta_{\mathbf{l}}$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\eta_{\mathbf{l}}$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \quad (3)$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i\left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(a_{l,d} - \eta_{\mathbf{l}}{}^T\bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d}y_{l,d} > 0\right). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics [18].

Posterior inference is performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \quad (6)$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \quad (7)$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha)$, $p(\alpha')$, $p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(2, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

3.2 Prediction

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of someone who has been hospitalized. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [citation]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text. Aside from prediction, one of the goals is to compare the sensitivity of predictions from the HSLDA model in comparison to the codes in a case where a test closer to ground truth is available. For this we will compare whether predictions for the ICD-9 code associated with anemia are better predicted by HSLDA or by the ICD-9 codes. Anemia was chosen because hemoglobin values are readily available and the definition of anemia according the World Health Organization is approximately 12.5, with a threshold of 12 for women and 13 for men [citation].

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset and partially directly from the the Amazon.com website. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.3 Evaluation

The two main methods of evaluation for the model are prediction and topic quality. We compare model performance against 3 similar models to demonstrate that each component of the model is important for performance. Specifically, we evaluate models including independent regressors + sLDA (hierarchical constraints on labels ignored), HSLDA fit by running LDA first then running tree-conditional regressions, and HSLDA fit with fixed random regression parameters.

4.3.1 Prediction

Given the expectation o

4.3.2 Topic Quality and Character

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models - predicting document links, other supervised latent variable models

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years.

A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

4.5 Evaluation

5 Results

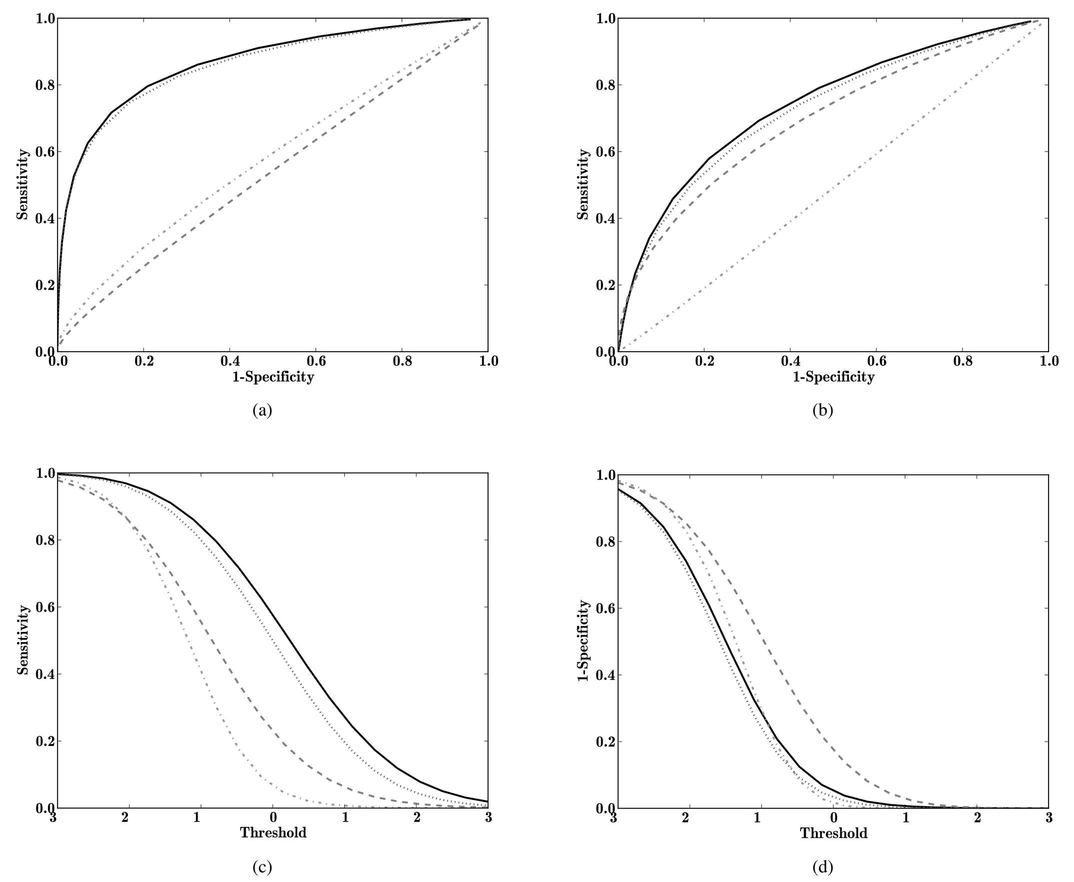


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

6 Discussion

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

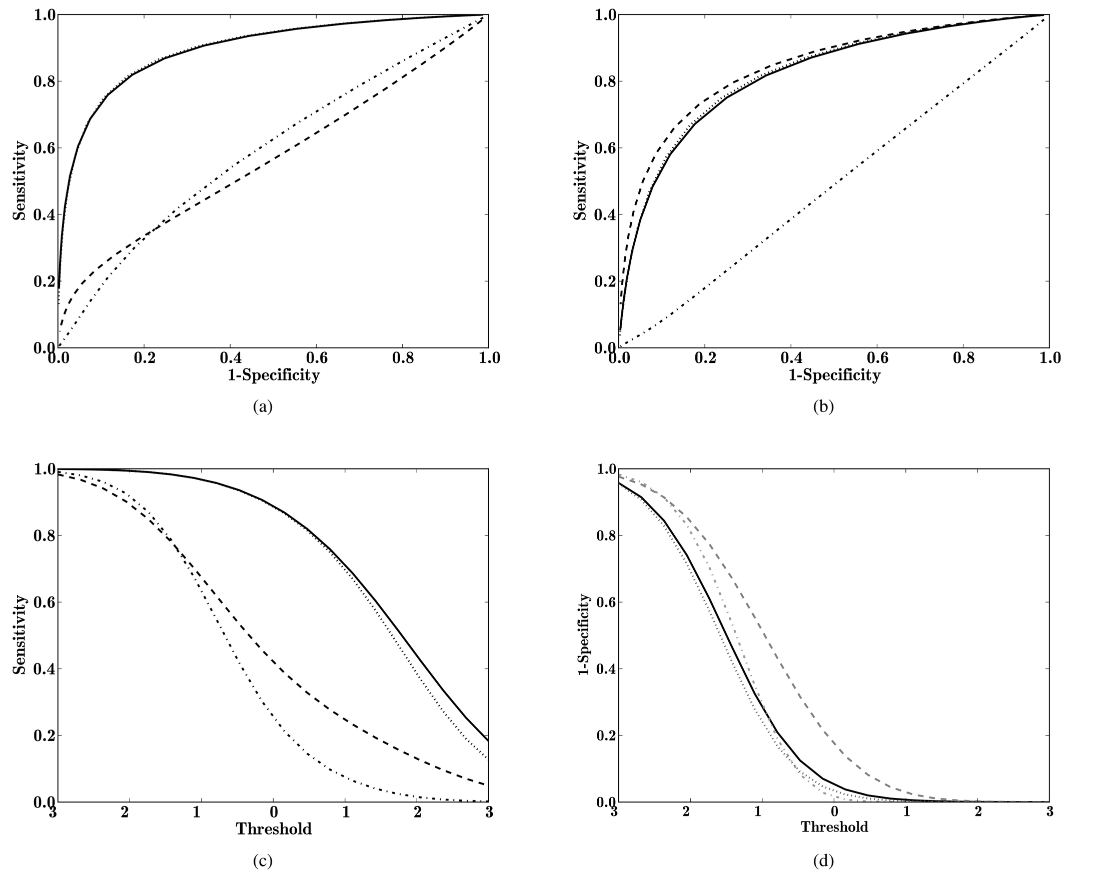


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] International classification of disease. http://bioportal.bioontology.org/ontologies/35686, May 2008.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

6.1 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks and show both improved label prediction performance and show evidence that the learned topic model improves as a result of using this signal too.

1 Introduction

There exist many models of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned []). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision;" often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditional draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchically constrained labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section ?? we introduce hierarchically supervised LDA (HSLDA), in Section 4.4 we review related work, and in Section ?? we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. We assume a pre-specified set of labels $\mathcal{L}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many, label(s) in $\mathcal{L}$. A response of 1 or -1 indicates if the document respectively is or is not l. The label l for a document d will be used interchangeably to refer the observed response of document d to label l. The label set is assumed to be an "is a" hierarchy. This means that if a label l_1 is a parent of label l_2 in the hierarchy and document d has a positive response to the label l_2 then document d also has a positive response to the label l_1. Seen in a negative light, if document d has a negative response to the l_1 label then document d also has a negative response to the l_2 label. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models. Approximate inference is performed using Gibbs sampling.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, and the standard deviation σ used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

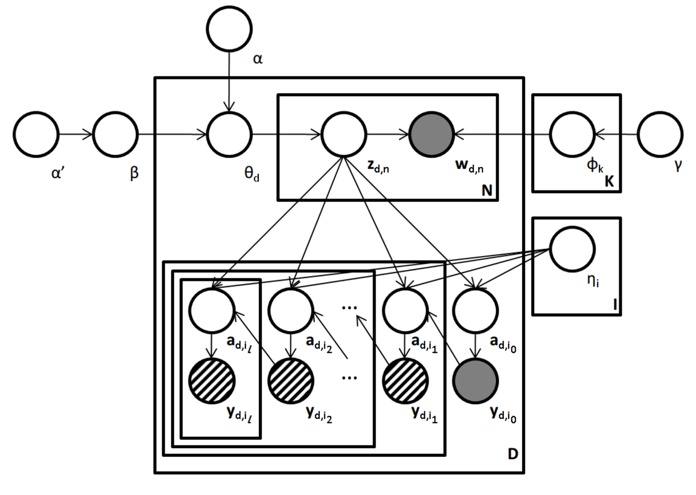


Figure 1: adapted sLDA model

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(-\mathbf{1}, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$:
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim \mathrm{Multinomial}(\beta_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l, y_{parent(l),d} \sim \begin{cases} \mathcal{N}(\bar{z}^T \eta_l, 1), & y_{parent(l)} = 1 \\ \mathcal{N}(\bar{z}^T \eta_l, 1)\, I(a_{l,d} < 0), & y_{parent(l)} = -1 \end{cases}$ where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{parent(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression is like logistic regression except instead of modeling the logit of $P(y = 1)$ using a linear form we model $\Phi^{-1}(P(y = 1))$ using a linear form, where $\Phi(\cdot)$ is the CDF for a standard normal distribution. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

3 Inference

In the Bayesian approach to statistical modeling the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ as in Griffiths and Steyvers [12]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\eta = \{\eta_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l,d}\}_{l \in \mathcal{L}, d=1,\ldots,D}$, β, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \setminus z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [12]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket, explicitly this means

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant [12] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients $\mathbf{eta_l}$ for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\begin{aligned} \hat{\Sigma}^{-1} &= \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} & (3) \\ \hat{\mu} &= \hat{\Sigma}_l \left(-\mathbf{1}\sigma^{-1} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). & (4) \end{aligned}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)^2\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion. This flexible distribution allows for an asymmetric prior over document level distributions over topics [18].

Posterior inference is performed using the "direct assignment" method of Teh et al. [17].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1}, m_{(\cdot),2}, \ldots, m_{(\cdot),K}\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(2, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of someone who has been hospitalized. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [citation]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text. Aside from prediction, one of the goals is to compare the sensitivity of predictions from the HSLDA model in comparison to the codes in a case where a test closer to ground truth is available. For this we will compare whether predictions for the ICD-9 code associated with anemia are better predicted by HSLDA or by the ICD-9 codes. Anemia was chosen because hemoglobin values are readily available and the definition of anemia according the World Health Organization is approximately 12.5, with a threshold of 12 for women and 13 for men [citation].

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset and partially directly from the the Amazon.com website. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.3 Evaluation

The two main methods of evaluation for the model are prediction and topic quality. We compare model performance against 3 similar models to demonstrate that each component of the model is important for performance. Specifically, we evaluate models including independent regressors + sLDA (hierarchical constraints on labels ignored), HSLDA fit by running LDA first then running tree-conditional regressions, and HSLDA fit with fixed random regression parameters.

4.3.1 Prediction

The two measures for predictive performance used here include the true positive rate and the false positive rate. We evaluate model performance on held out data. A more ideal evaluation of performance would include a manually labeled hierarchy since it is well known that ICD-9 codes have a relatively low sensitivity.

Performance was evaluated against $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$.

4.3.2 Topic Quality and Character

TBD

4.4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In

other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [citations]. One set of models that are particularly relevant to HSLDA are Chang and Blei's [citation] hierarchical models for document networks. In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In our model, the lack of a code in the hierarchy being assigned does not necessarily indicate absence. Therefore, as in the work of Chang and Blei, we employ regularization in the form of a negative prior on the regression parameters to provide a prior that indicates a bias towards being truly negative in the absence of a code.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [16, 10, 15, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [14] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [11, 8, 9]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [13]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[14] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[15] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[16] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

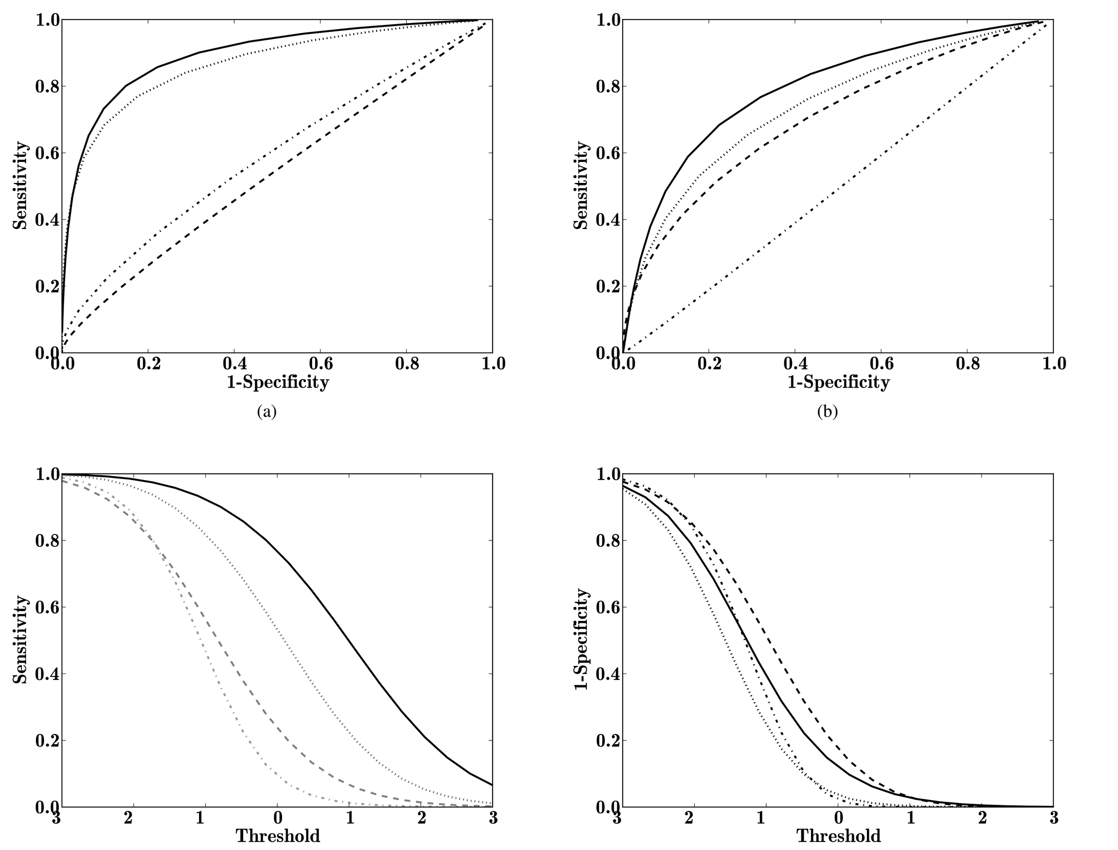


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

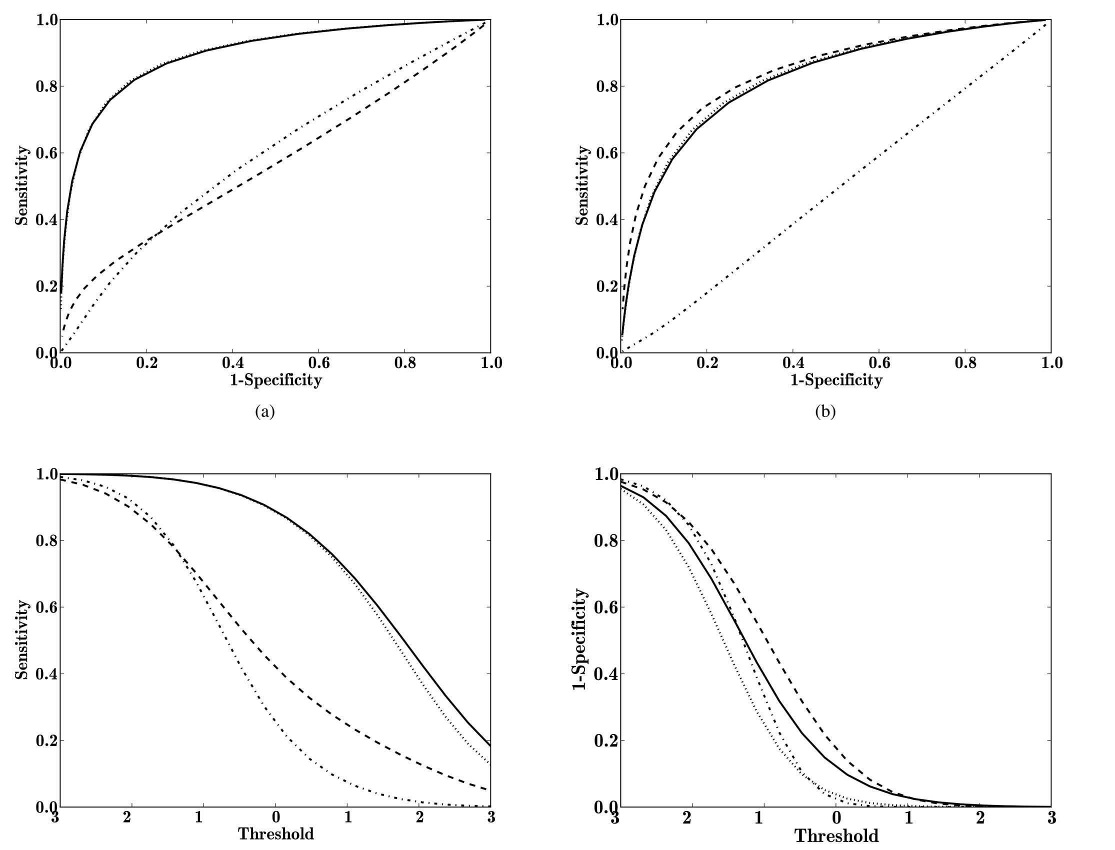


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [3]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [**?**]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4.4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section **??** we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

There have been many models that incorporate both latent models of text and some form of supervision [**? ? ? ?**]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [15, 9, 14, 6], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [13] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [10, 7, 8]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [12]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

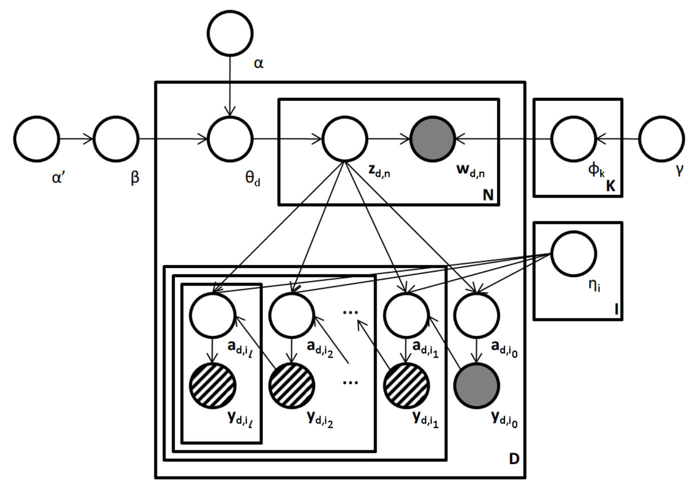


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(\mu I_K, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$

- For $n = 1, \ldots, N_d$:
  - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
  - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
- For each label $l \in \mathcal{L}$:
  - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l, y_{parent(l),d} \sim \begin{cases} \mathcal{N}(\bar{z}^T \eta_l, 1), & y_{parent(l)} = 1 \\ \mathcal{N}(\bar{z}^T \eta_l, 1)\, I(a_{l,d} < 0), & y_{parent(l)} = -1 \end{cases}$ where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
  - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{and } y_{parent(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\eta_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate out certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ as in Griffiths and Steyvers [11]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\eta = \{\eta_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, β, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \setminus z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [11]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [11] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d'}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients η_l for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l{}^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [17]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [16].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha)$, $p(\alpha')$, $p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [**?**]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [3] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first method augmented the observed labels in the held out set as well as their ancestors in the hierarchy as

4.3.1 Prediction

The two measures for predictive performance used here include the true positive rate and the false positive rate. We evaluate model performance on held out data. A more ideal evaluation of performance would include a manually labeled hierarchy since it is well known that ICD-9 codes have a relatively low sensitivity.

Performance was evaluated against $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$.

4.3.2 Topic Quality and Character

TBD

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[9] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[10] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[11] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[12] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[13] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[14] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[15] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[17] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

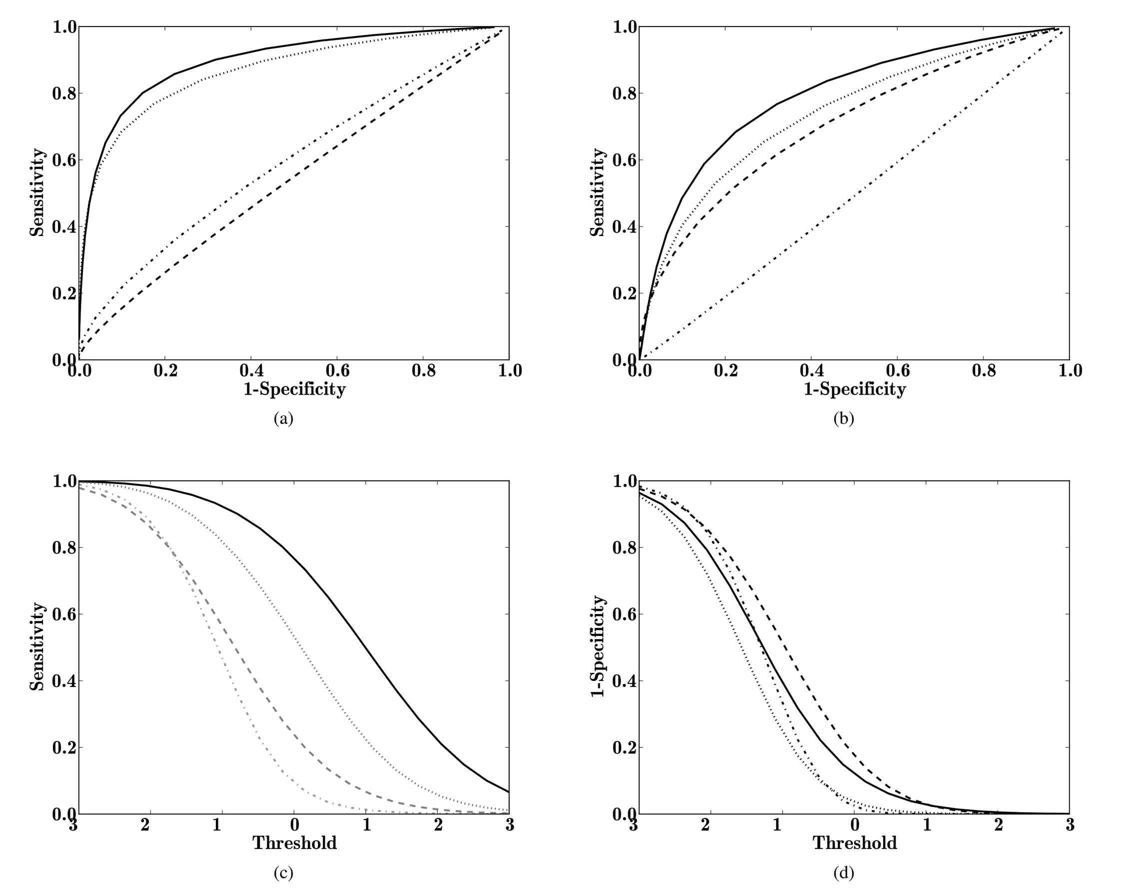


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

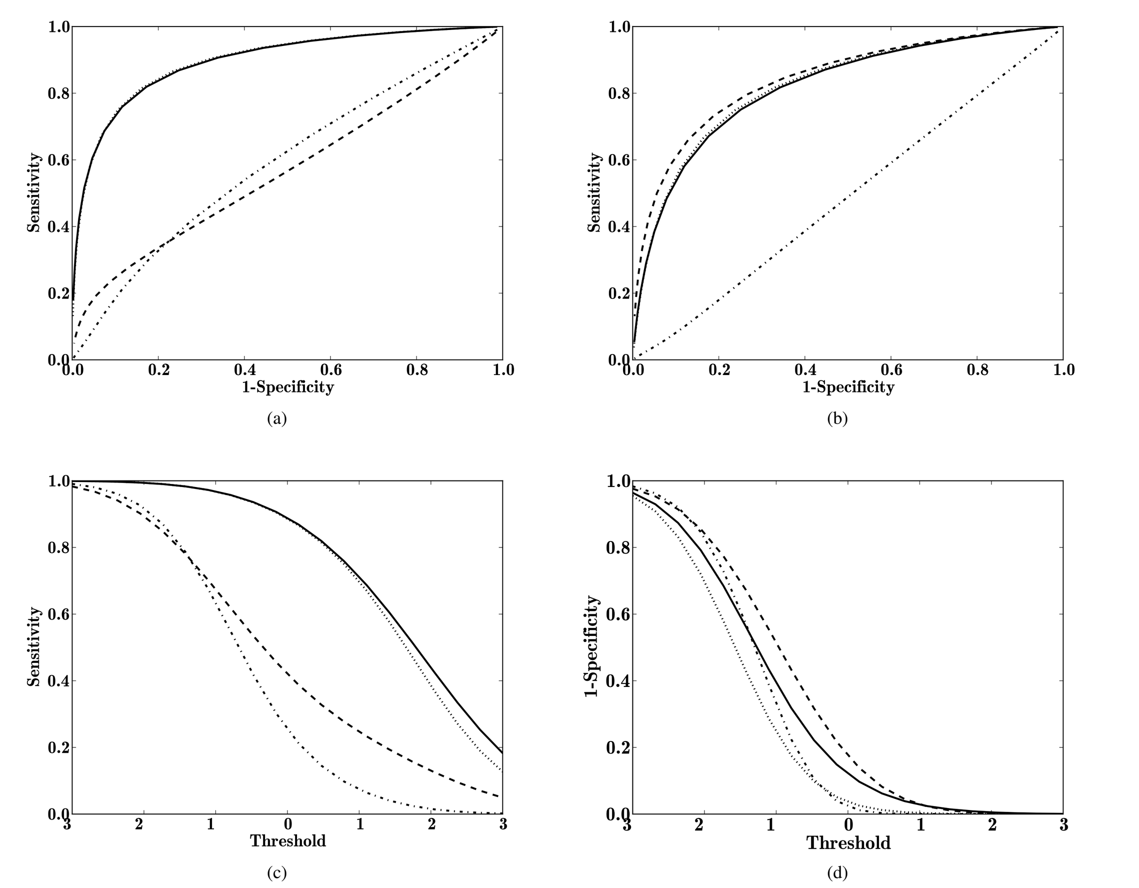


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [3]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [?]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

There have been many models that incorporate both latent models of text and some form of supervision [? ? ? ?]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [15, 9, 14, 6], fully automatic assignment of ICD-9 codes to medical text has become a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [13] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [10, 7, 8]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [12]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

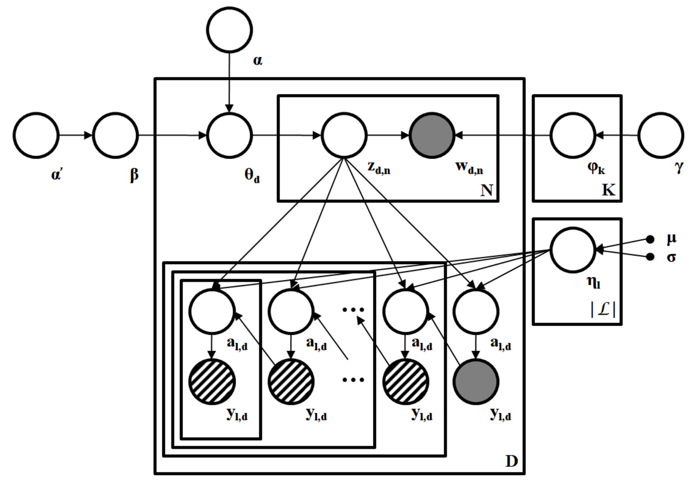


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(\mu I_K, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$:
     - Draw topic assignment $z_{n,d} \mid \theta_\mathbf{d} \sim \mathrm{Multinomial}(\theta_\mathbf{d})$
     - Draw word $w_{n,d} \mid z_{n,d}, \beta_{\mathbf{1:K}} \sim \mathrm{Multinomial}(\beta_{\mathbf{z_{n,d}}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l, y_{parent(l),d} \sim \begin{cases} \mathcal{N}(\bar{z}^T\eta_l, 1), & y_{parent(l)} = 1 \\ \mathcal{N}(\bar{z}^T\eta_l, 1)\, I(a_{l,d} < 0), & y_{parent(l)} = -1 \end{cases}$ where $\bar{z}_d = N_d^{-1}\sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{and } y_{\text{parent}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\eta_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L}\backslash\mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [11]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \eta = \{\eta_\mathbf{l}\}_{l\in\mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l'\in\mathcal{L}_d, d=1,\ldots,D}, \beta, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}\backslash z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [11]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l\in\mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [11] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l\in\mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T\eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_\mathbf{l} \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients $\eta_\mathbf{l}$ for $l \in \mathcal{L}$. Given that $\eta_\mathbf{l}$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\eta_\mathbf{l}$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\mathbf{\Sigma}}$ such that

$$\hat{\mathbf{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\mathbf{\Sigma}}_i\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_\mathbf{l}{}^T\bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d}y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [17]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [16].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha)$, $p(\alpha')$, $p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [?]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [3] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[8] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[9] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[10] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[11] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[12] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

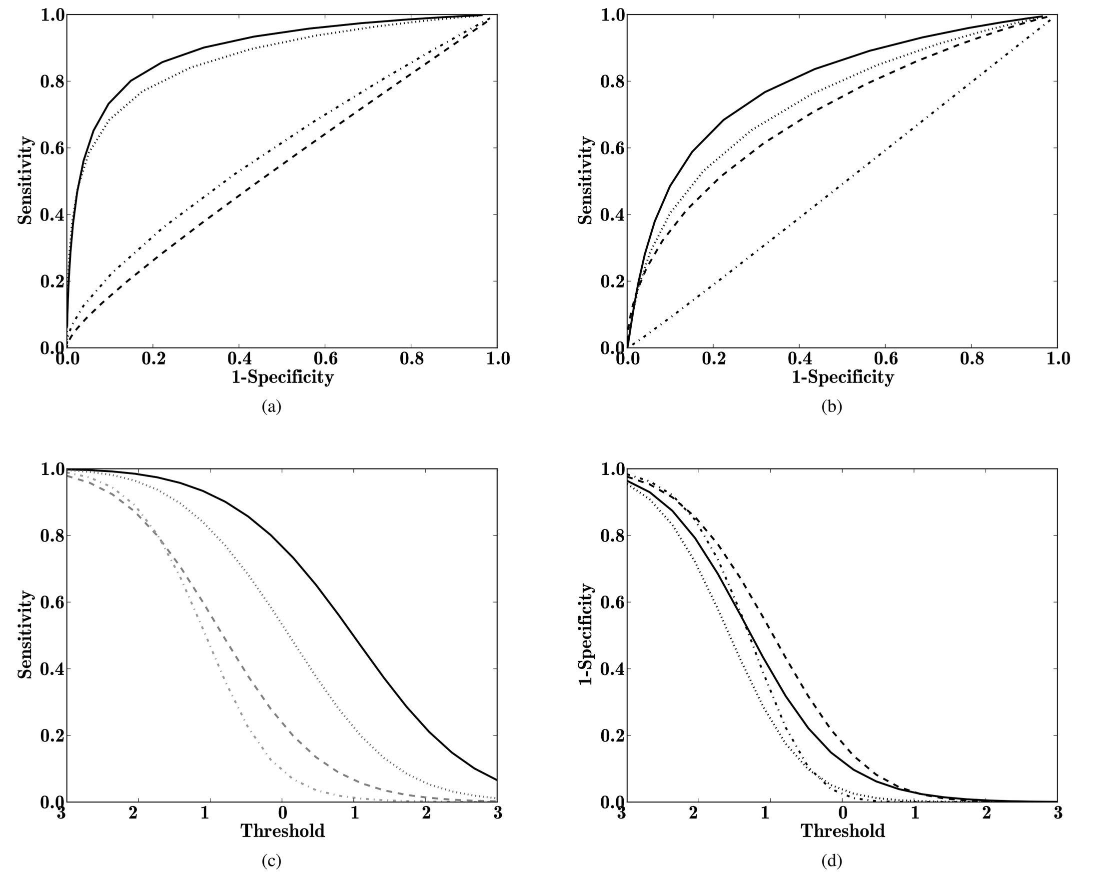


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

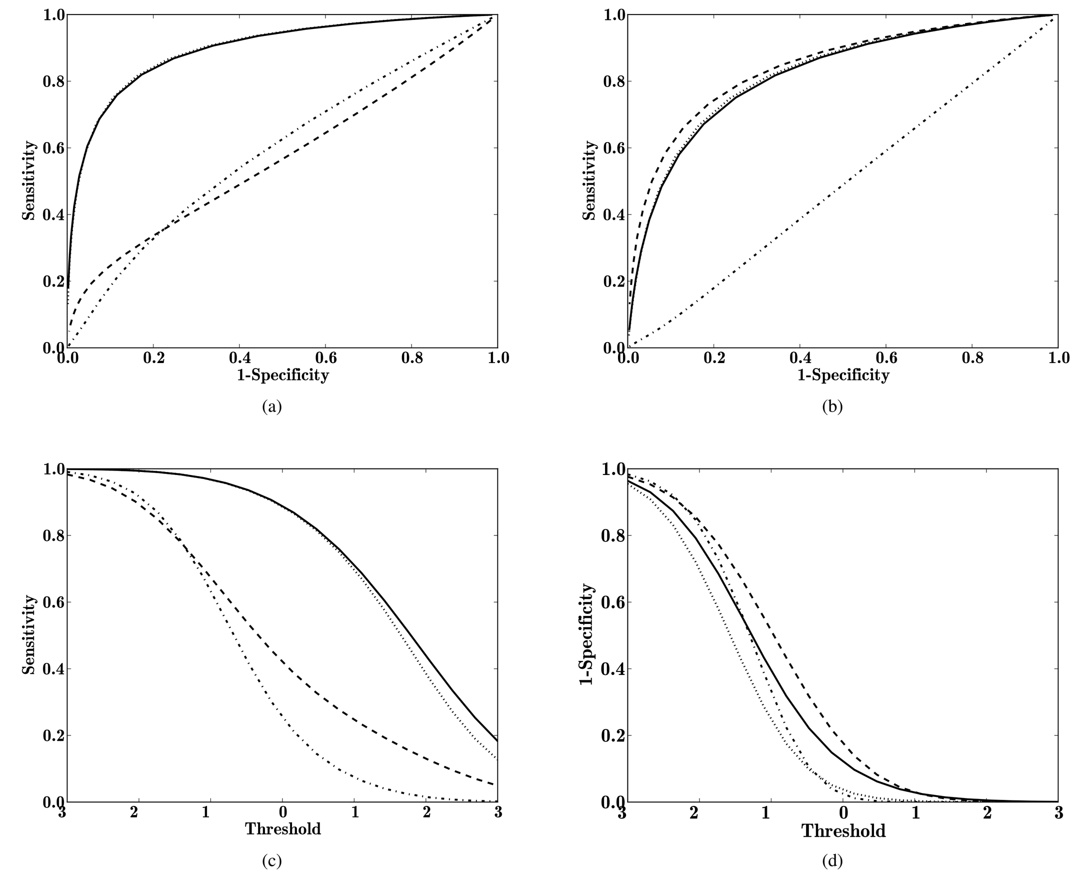


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[13] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[14] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[15] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[17] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

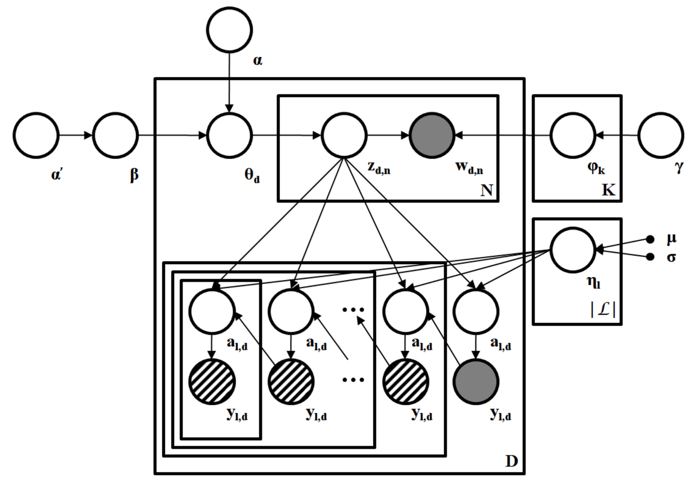


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$:
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1})$, where $\mathbf{1}$ is a vector of ones of length V
2. For each label $l \in \mathcal{L}$:
   - Draw the regression coefficients $\eta_l \mid \sigma \sim \mathcal{N}_K(\mu I_K, \sigma I_K)$, where I_K is the K dimensional identity matrix
3. Draw a prior over topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1})$
4. For each document $d = 1, \ldots, D$:
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$

- For $n = 1, \ldots, N_d$:
  - Draw topic assignment $z_{n,d} \mid \theta_{\mathbf{d}} \sim \mathrm{Multinomial}(\theta_{\mathbf{d}})$
  - Draw word $w_{n,d} \mid z_{n,d}, \beta_{\mathbf{1:K}} \sim \mathrm{Multinomial}(\beta_{\mathbf{z_{n,d}}})$
- For each label $l \in \mathcal{L}$:
  - Draw $a_{l,d} \mid z_{1:N_d,d}, \eta_l, y_{parent(l),d} \sim \begin{cases} \mathcal{N}(\bar{z}^T \eta_l, 1), & y_{parent(l)} = 1 \\ \mathcal{N}(\bar{z}^T \eta_l, 1)\, I(a_{l,d} < 0), & y_{parent(l)} = -1 \end{cases}$ where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
  - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{and } y_{\mathrm{parent}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\eta_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \eta = \{\eta_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \beta, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \backslash z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{n^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(n^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients η_l for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\mathbf{\Sigma}}$ such that

$$\begin{aligned} \hat{\mathbf{\Sigma}}^{-1} &= \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3) \\ \hat{\mu} &= \hat{\mathbf{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4) \end{aligned}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l{}^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{7}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

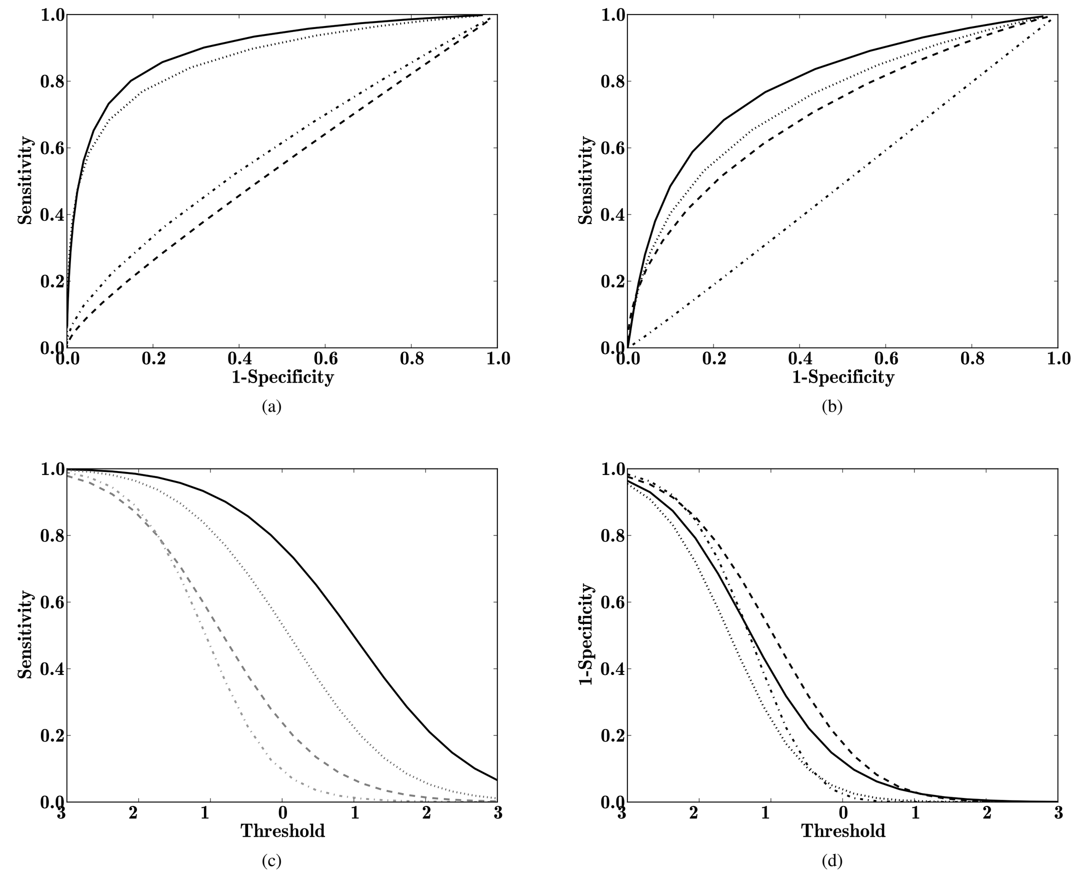


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

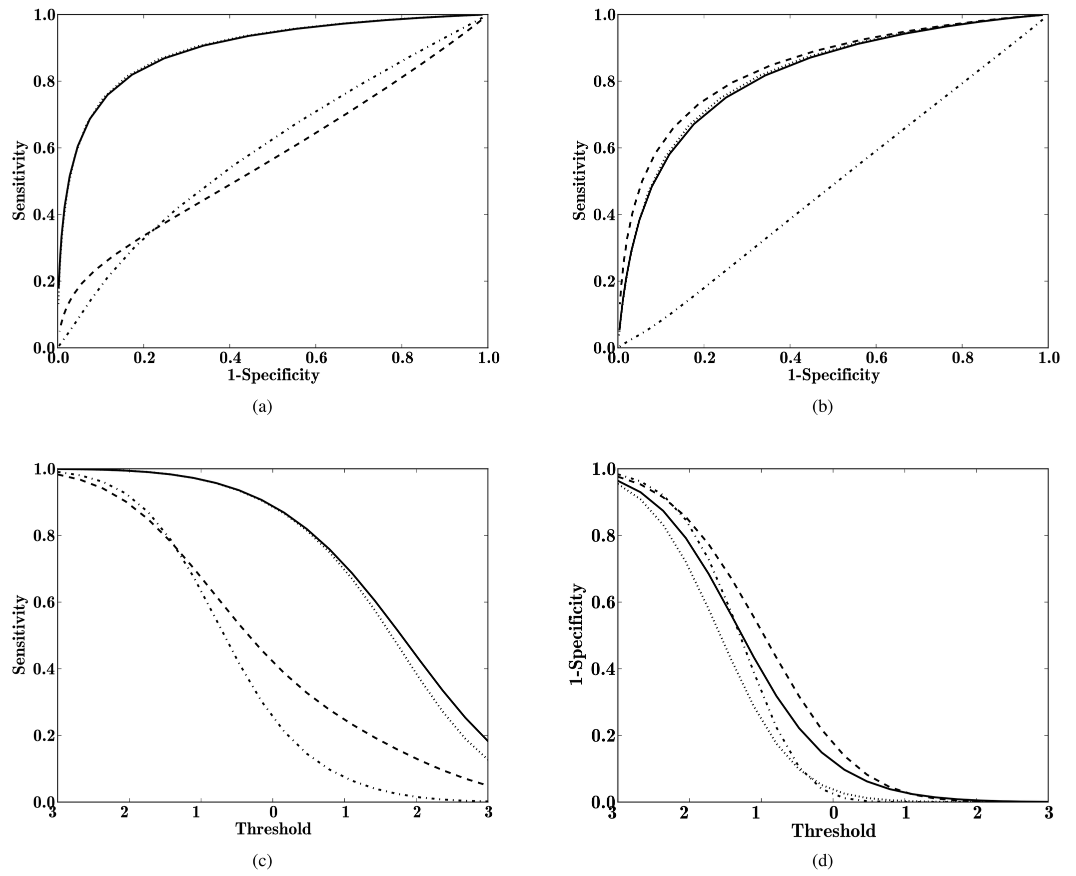


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressions on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchical supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hiearchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

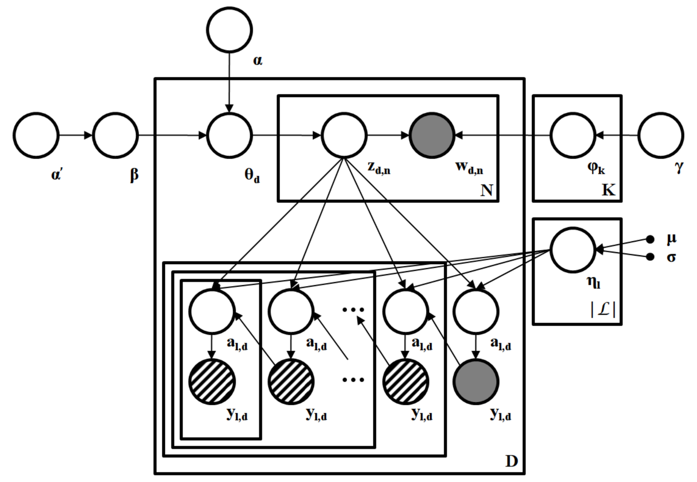


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$

- For $n = 1, \ldots, N_d$
  - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
  - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
- For each label $l \in \mathcal{L}$:
  - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

    where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
  - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \eta = \{\eta_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \beta, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \setminus z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \prod_{l \in \mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \eta_l)\, p(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d'}^{k,-(n,d)}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients η_l for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l{}^T \bar{\mathbf{z}}_d\right)\right\} I(a_{l,d} y_{l,d} > 0). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\beta \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m)(\alpha\beta_k)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

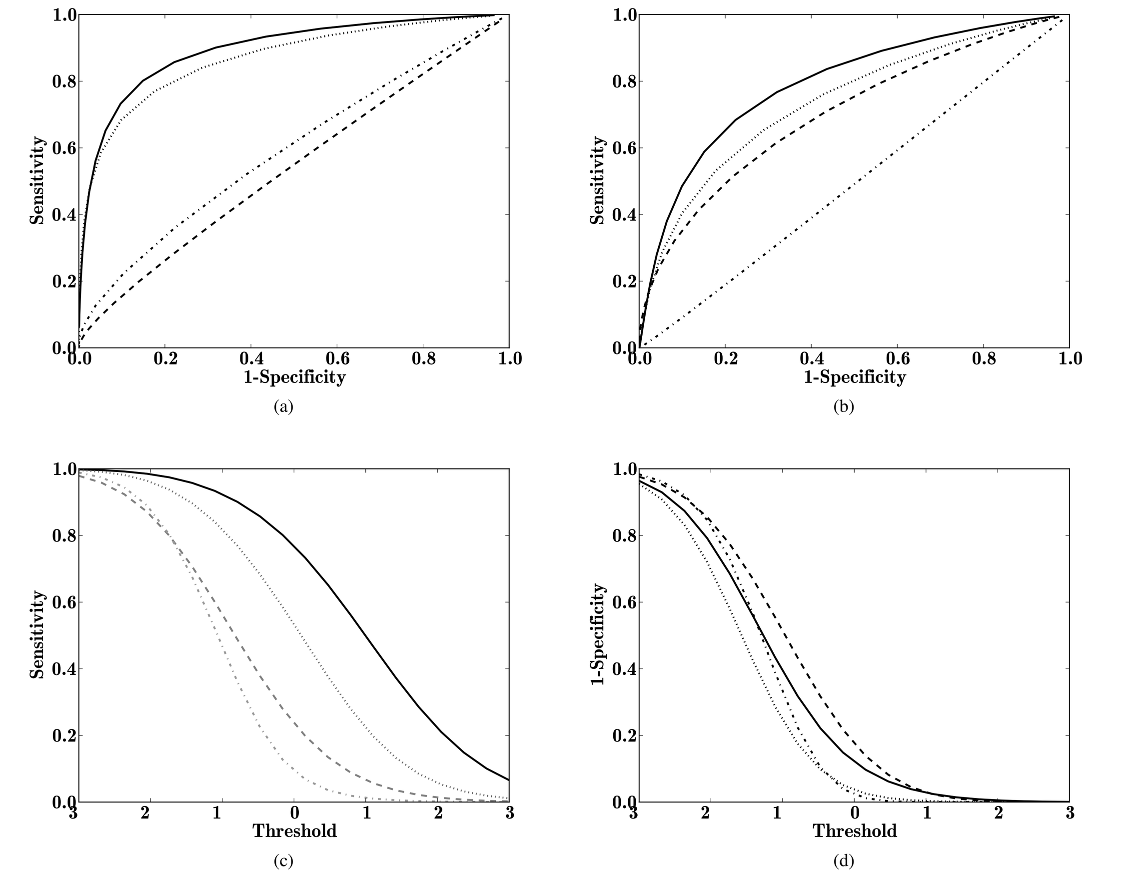


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

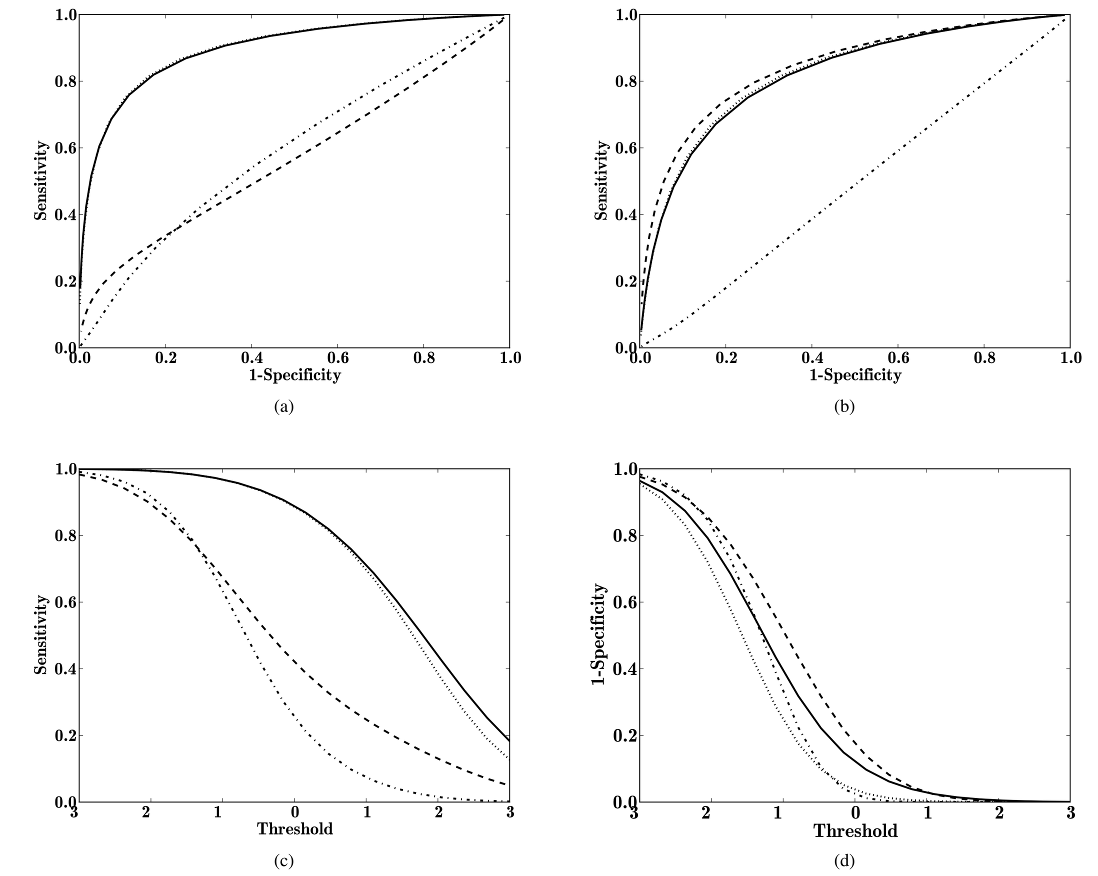


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchical supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

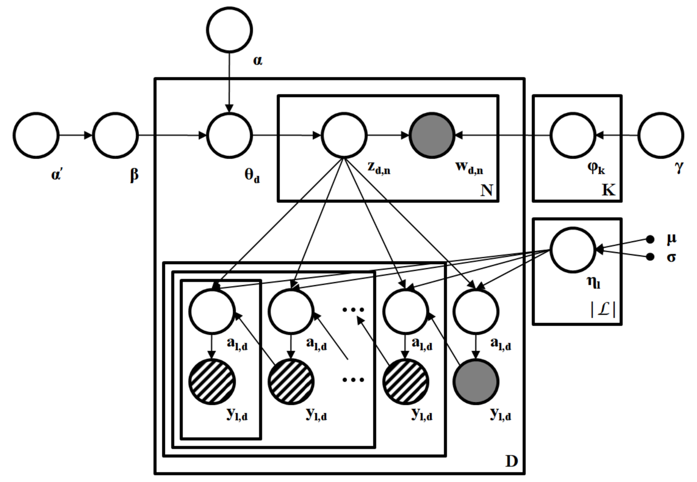


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$

- For $n = 1, \ldots, N_d$
  - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
  - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
- For each label $l \in \mathcal{L}$:
  - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

    where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
  - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \eta = \{\eta_\mathbf{l}\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \beta, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \backslash z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto$$
$$\left(n^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{n^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(n^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_\mathbf{l} \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients $\eta_\mathbf{l}$ for $l \in \mathcal{L}$. Given that $\eta_\mathbf{l}$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\eta_\mathbf{l}$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\mathbf{\Sigma}}$ such that

$$\hat{\mathbf{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$
$$\hat{\mu} = \hat{\mathbf{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_\mathbf{l}{}^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

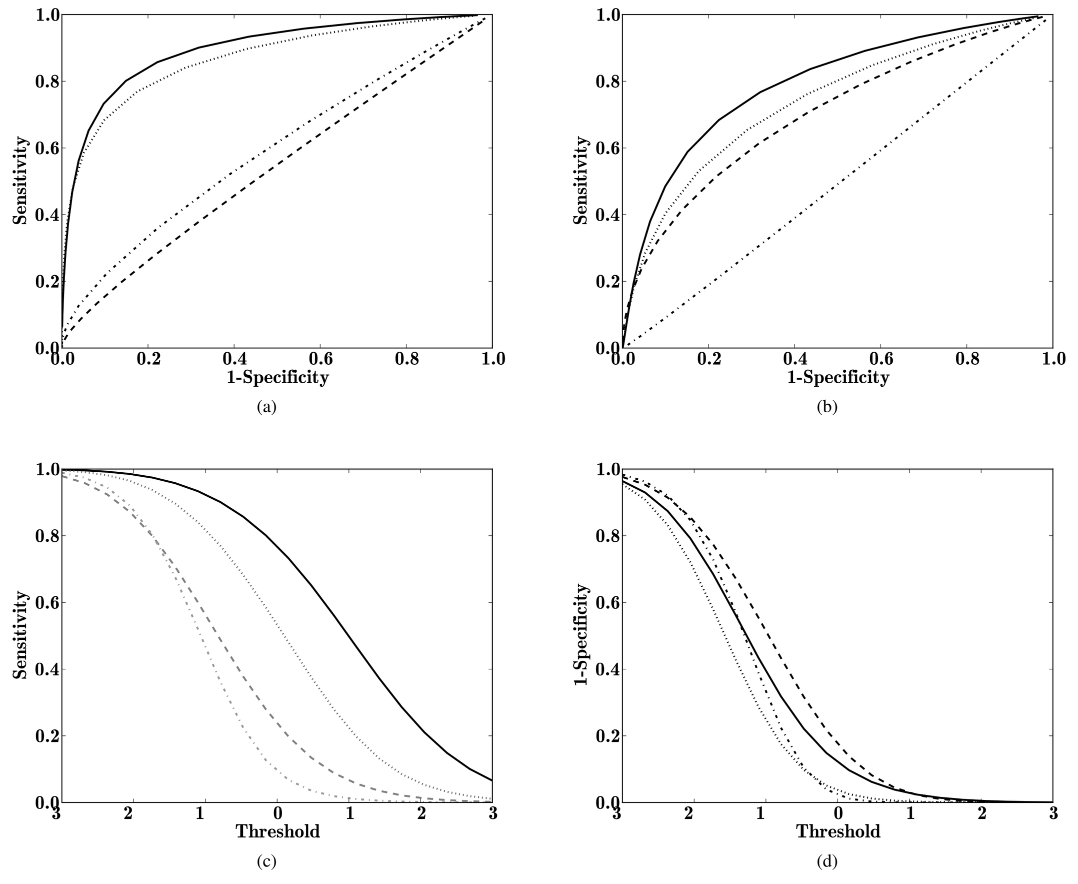


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

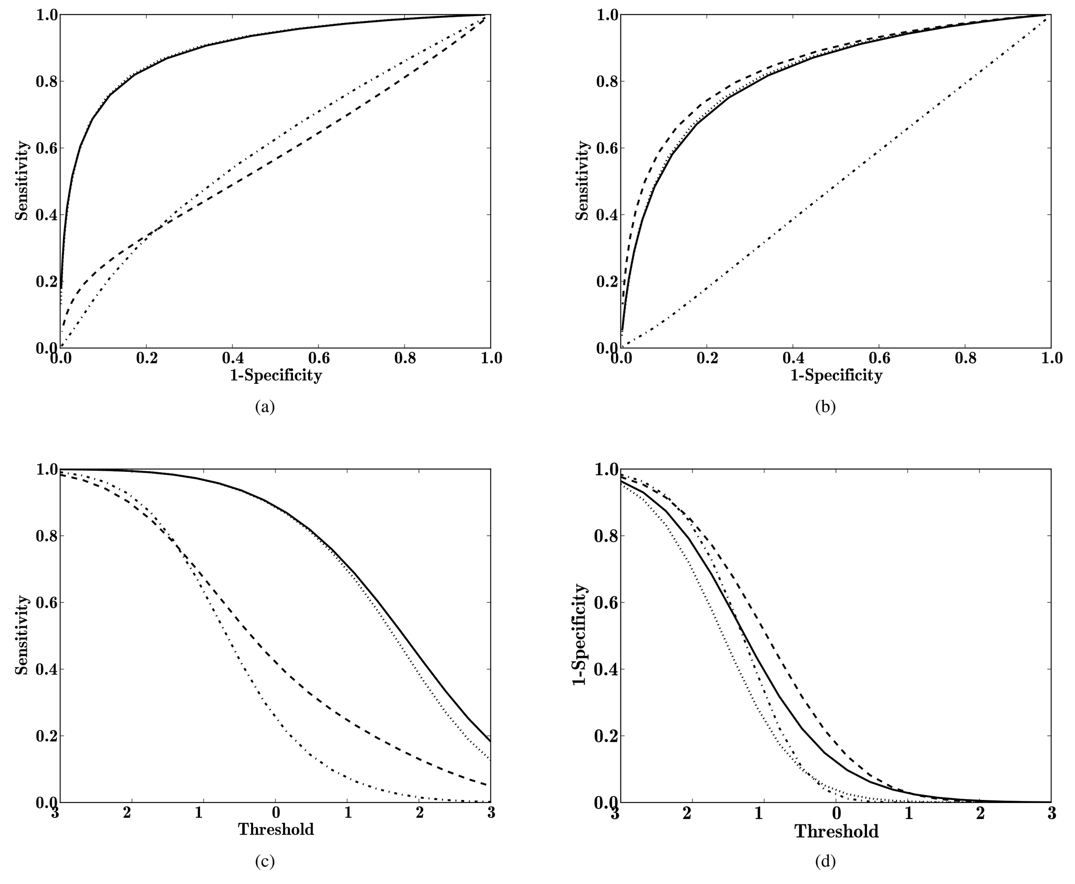


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hieararchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

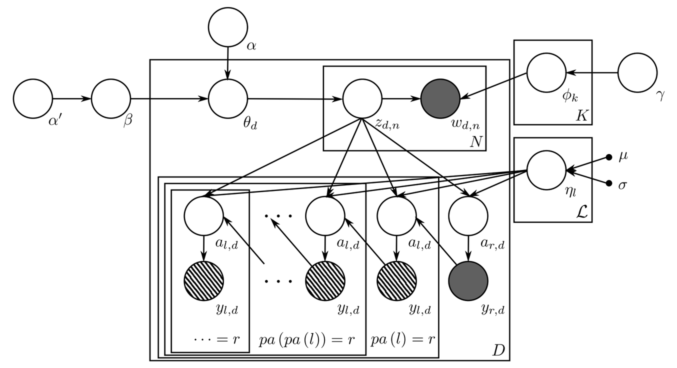


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \dots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \dots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \dots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{pa(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1), & y_{pa(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{pa(l),d} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{pa(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{i,N_d,d}\}_{d=1,\dots,D}, \eta = \{\eta_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\dots,D}, \beta, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z} \backslash z_{n,d}$.

3.1.1 $p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \dots, N_d$ in documents $d = 1, \dots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \eta_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \alpha, \beta, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(\mathbf{n},\mathbf{d})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{z}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p\left(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma\right)$

We now consider the conditional distribution of the regression coefficients η_l for $l \in \mathcal{L}$. Given that η_l and $a_{l,d}$ are distributed normally, the posterior distribution of η_l is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{z}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \dots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \dots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l{}^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\beta \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\beta \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \dots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\beta_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \dots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

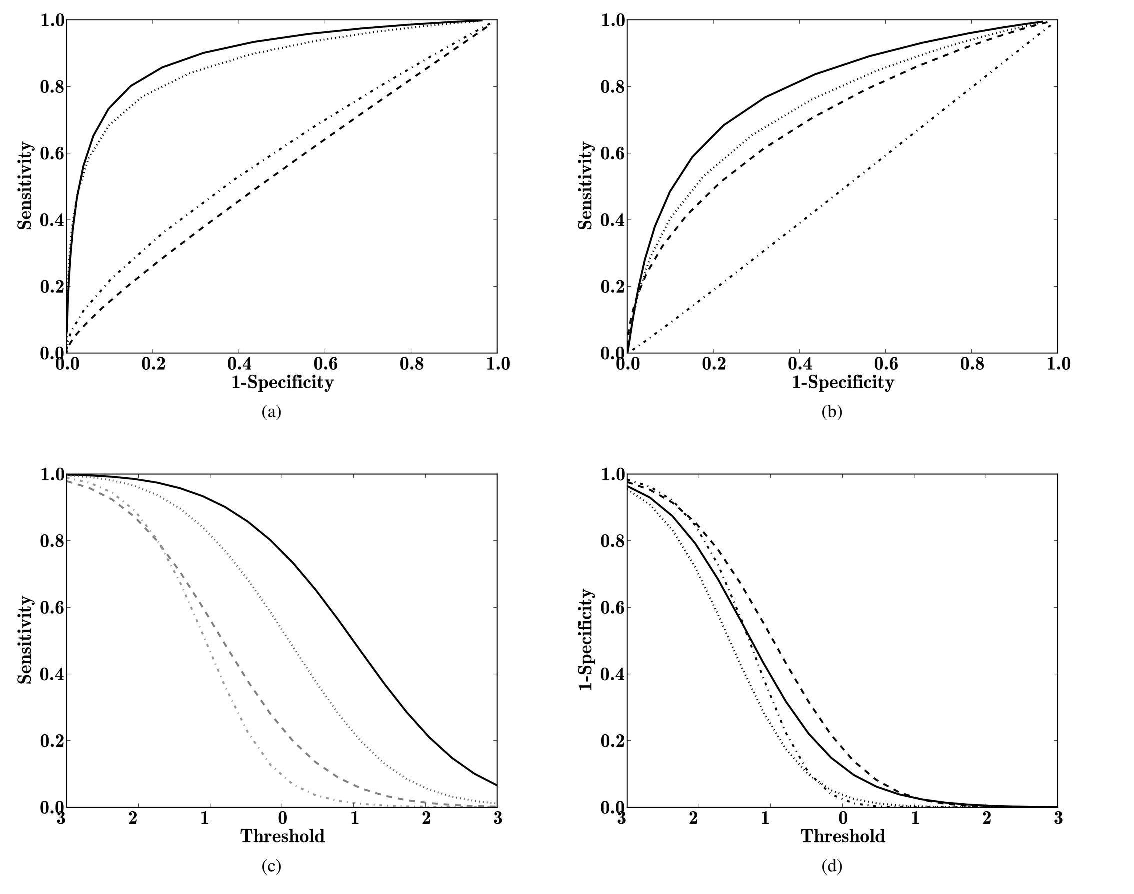


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

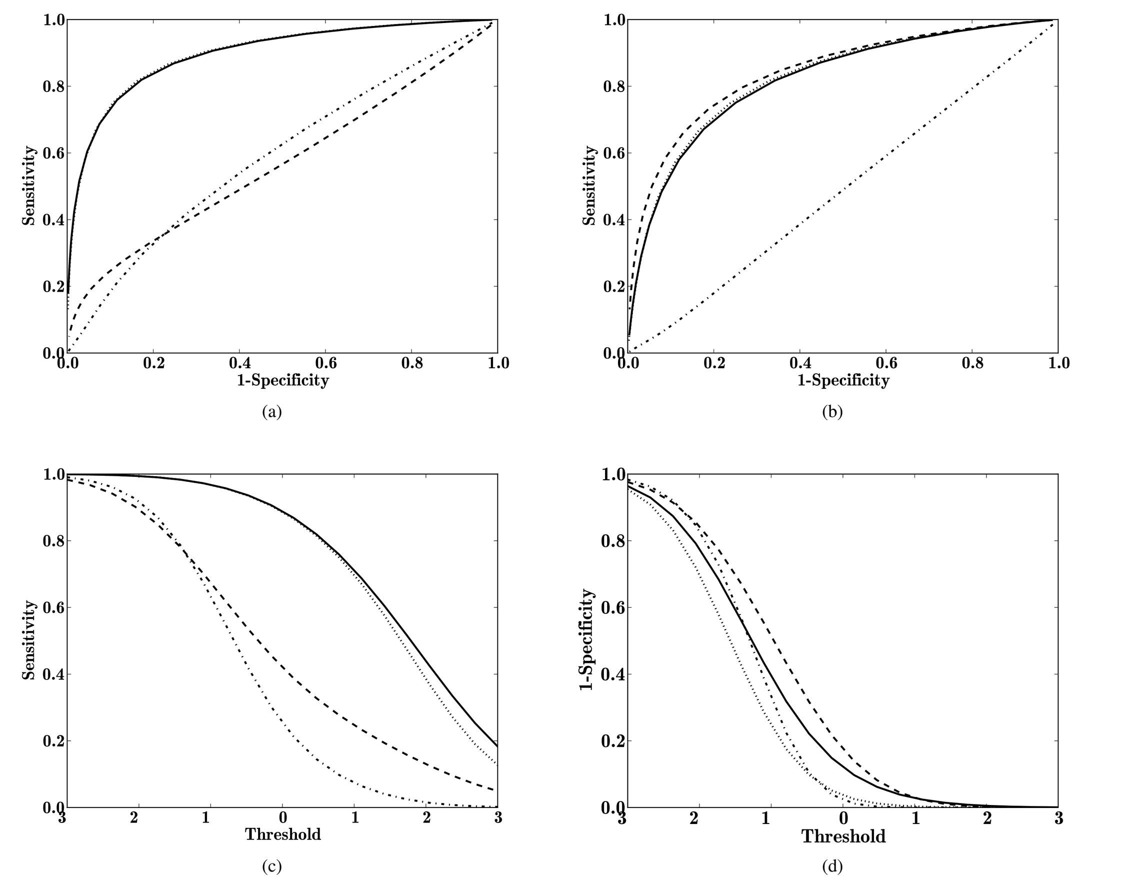


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte   Nicholas Bartlett   Noemie Elhadad   Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

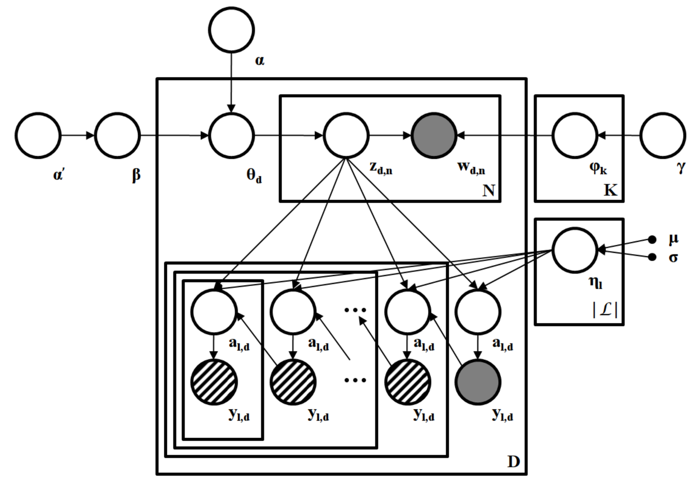


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$

   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$

$$\left(n_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

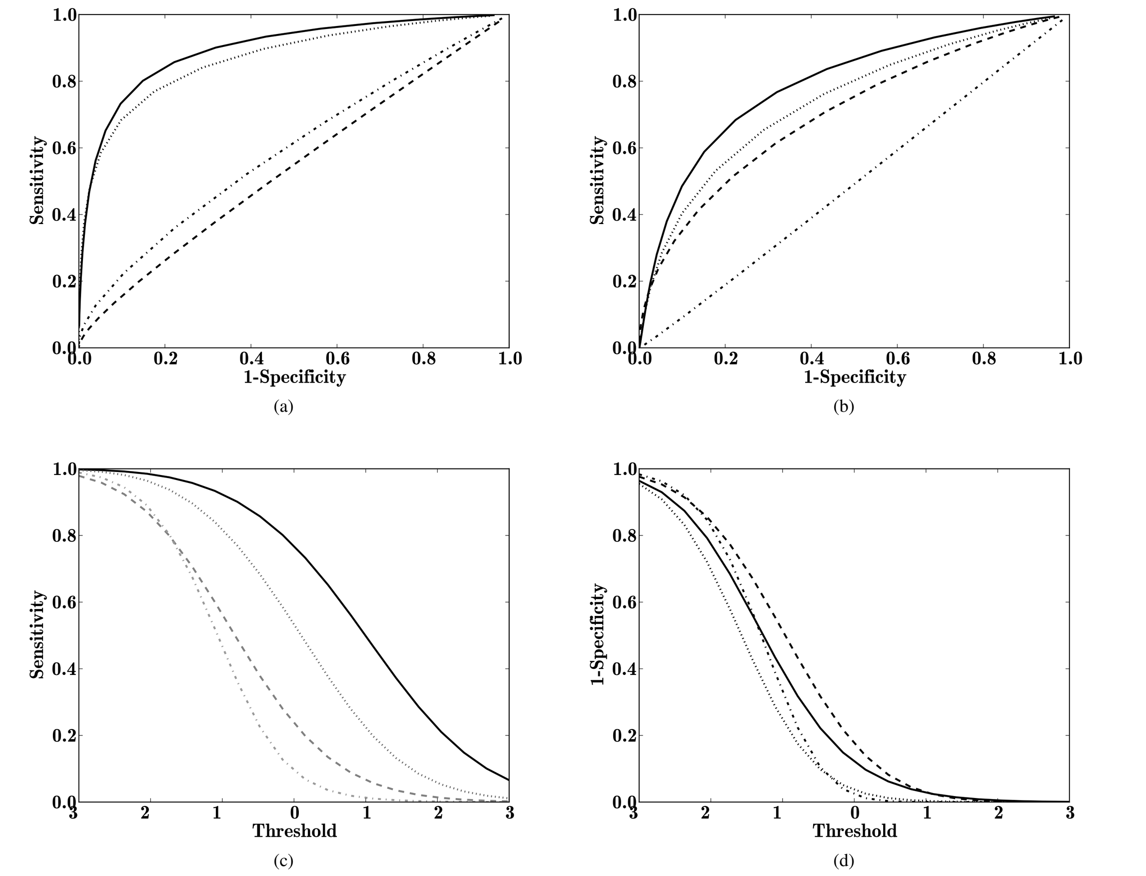


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

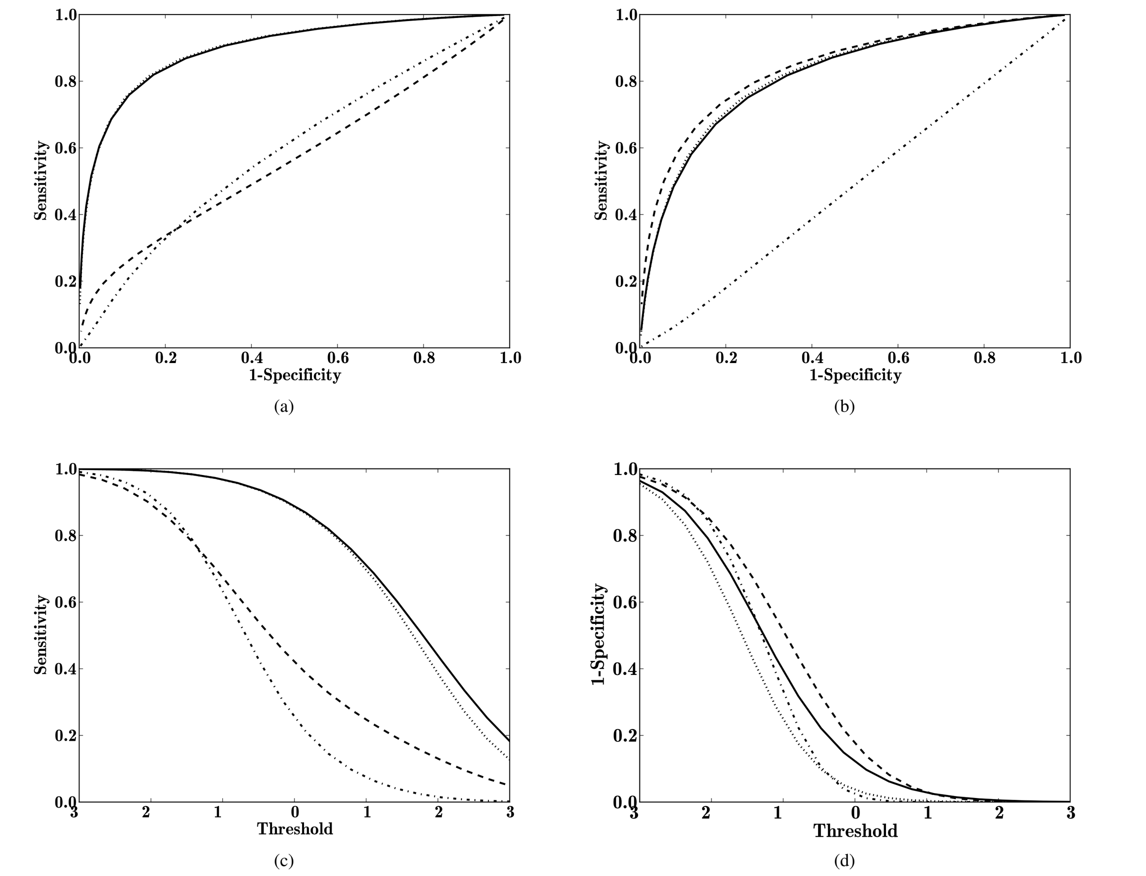


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte   Nicholas Bartlett   Noemie Elhadad   Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

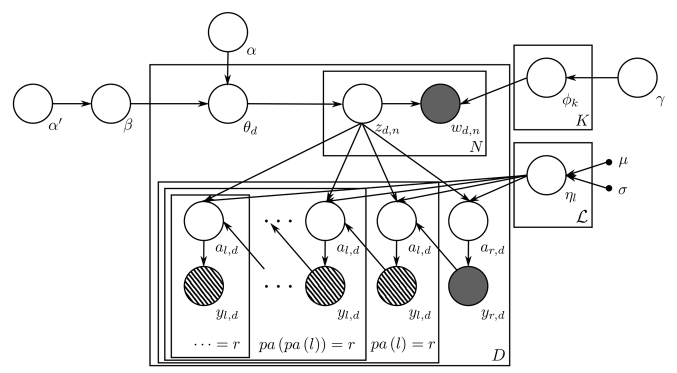


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     – Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     – Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     – Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\beta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     – Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod\nolimits_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right)$, $p\left(\alpha'\right)$, $p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

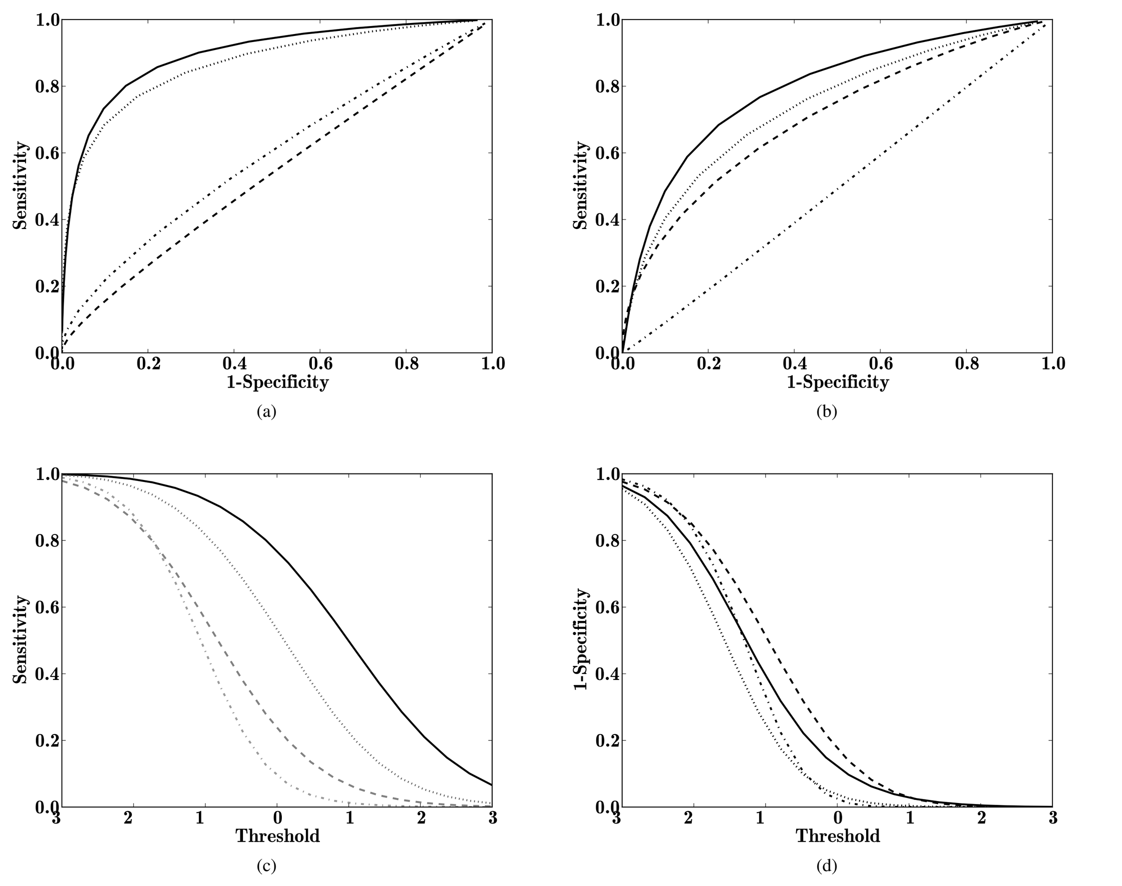


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

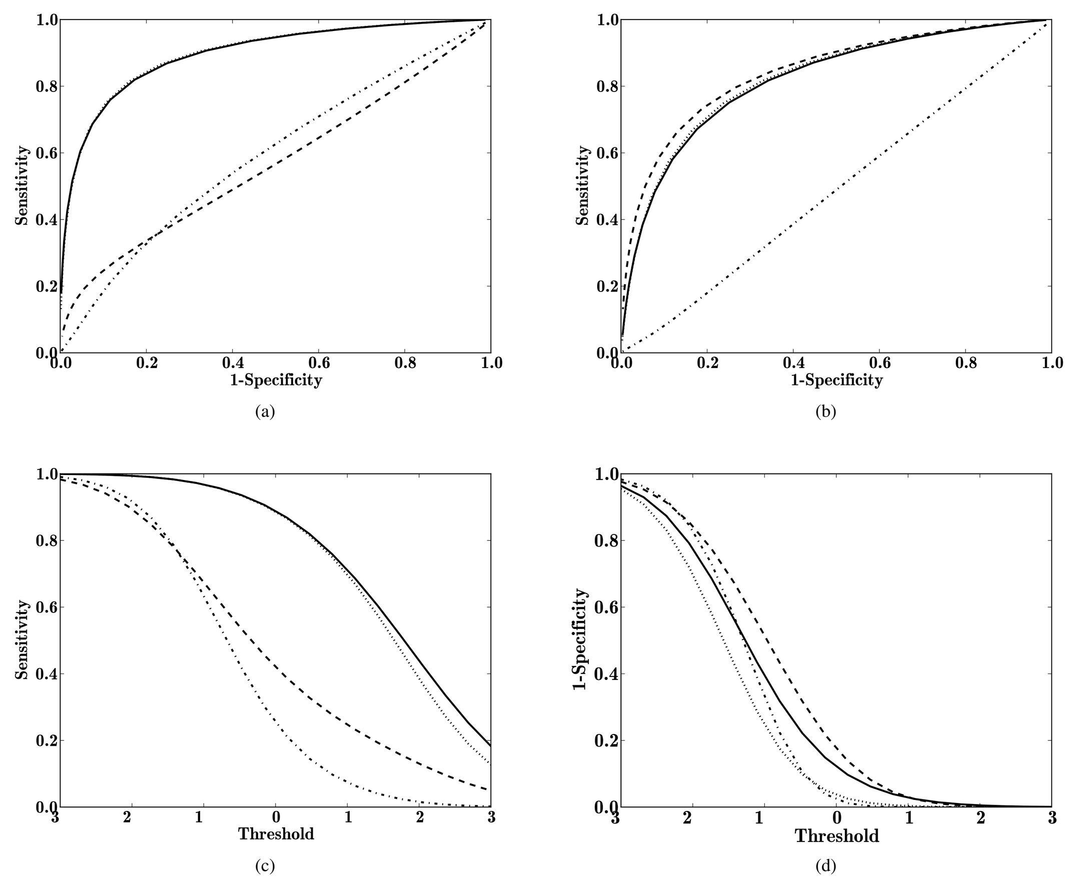


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressions on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hieararchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

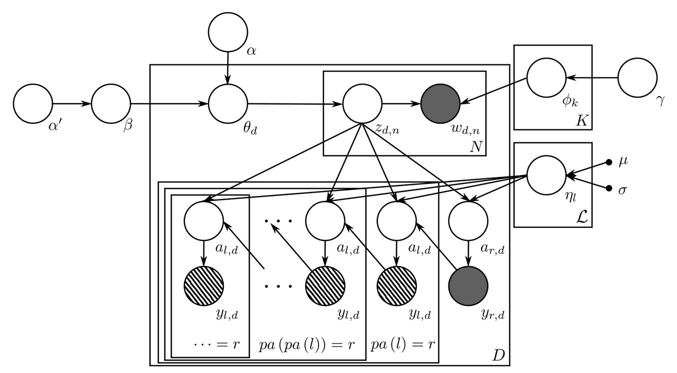


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} I\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

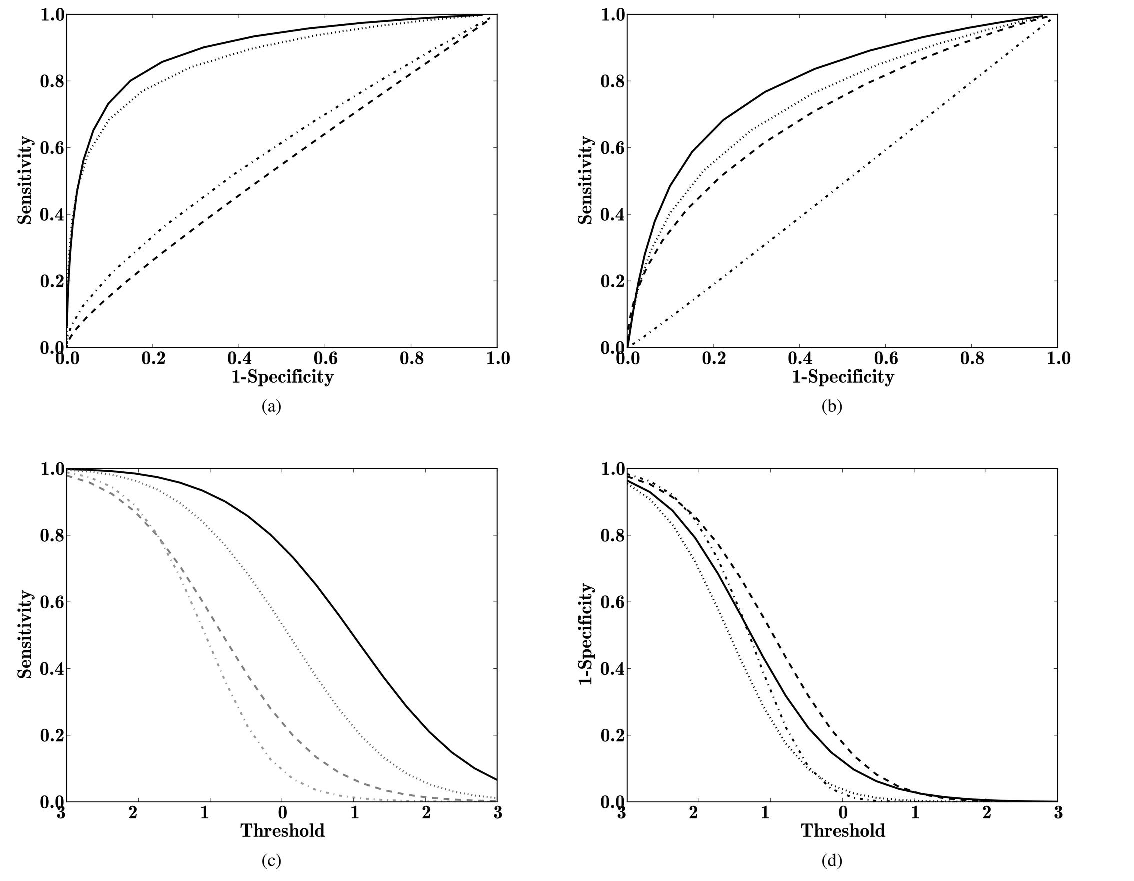


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

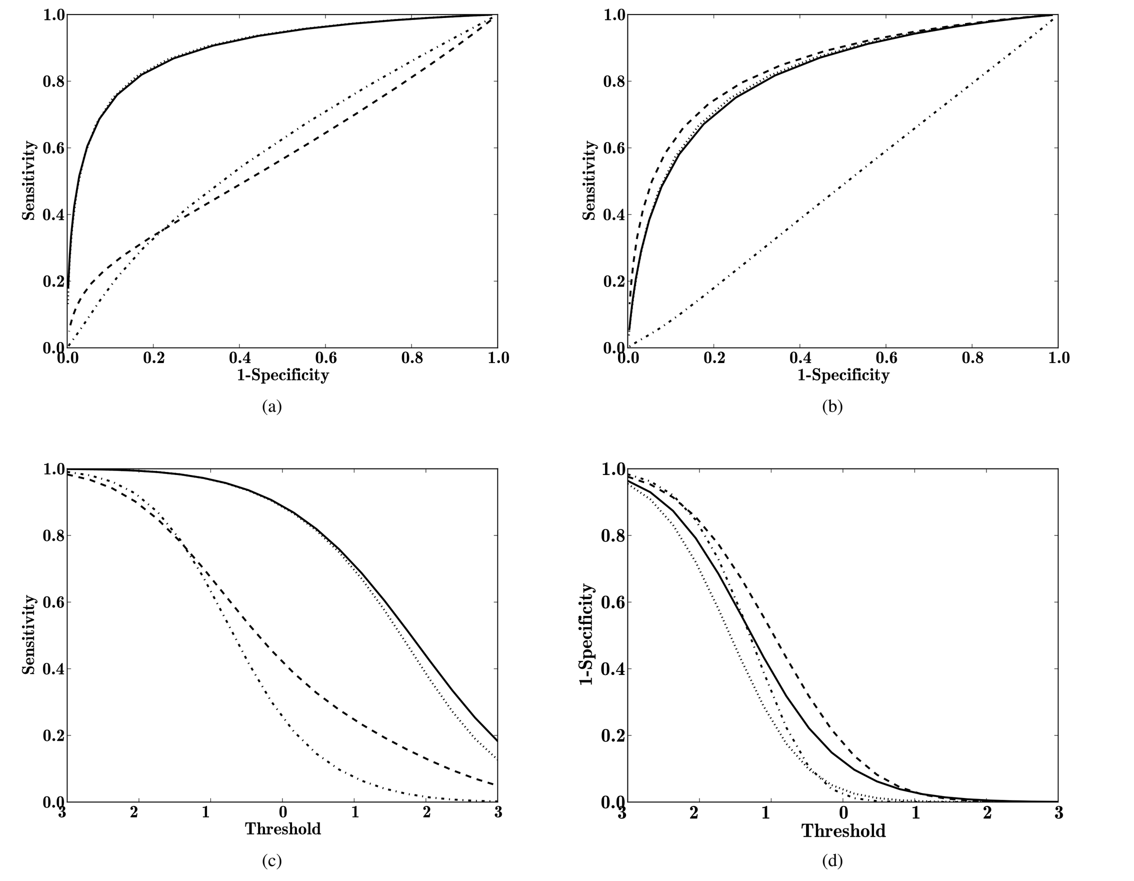


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hieararchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

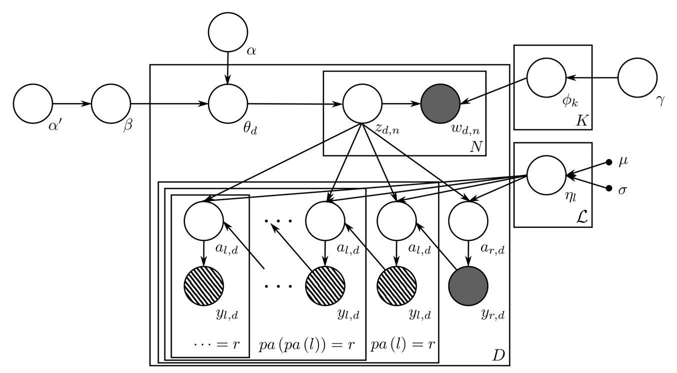


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{pa(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1), & y_{pa(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\beta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{pa(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{pa(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \setminus z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \prod_{l \in \mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l)\, p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\}. \quad (2)$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3)$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (6)$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma(\alpha \boldsymbol{\beta}_k + n_{d,k})} s(n_{d,k}, m) (\alpha \boldsymbol{\beta}_k)^m \quad (7)$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

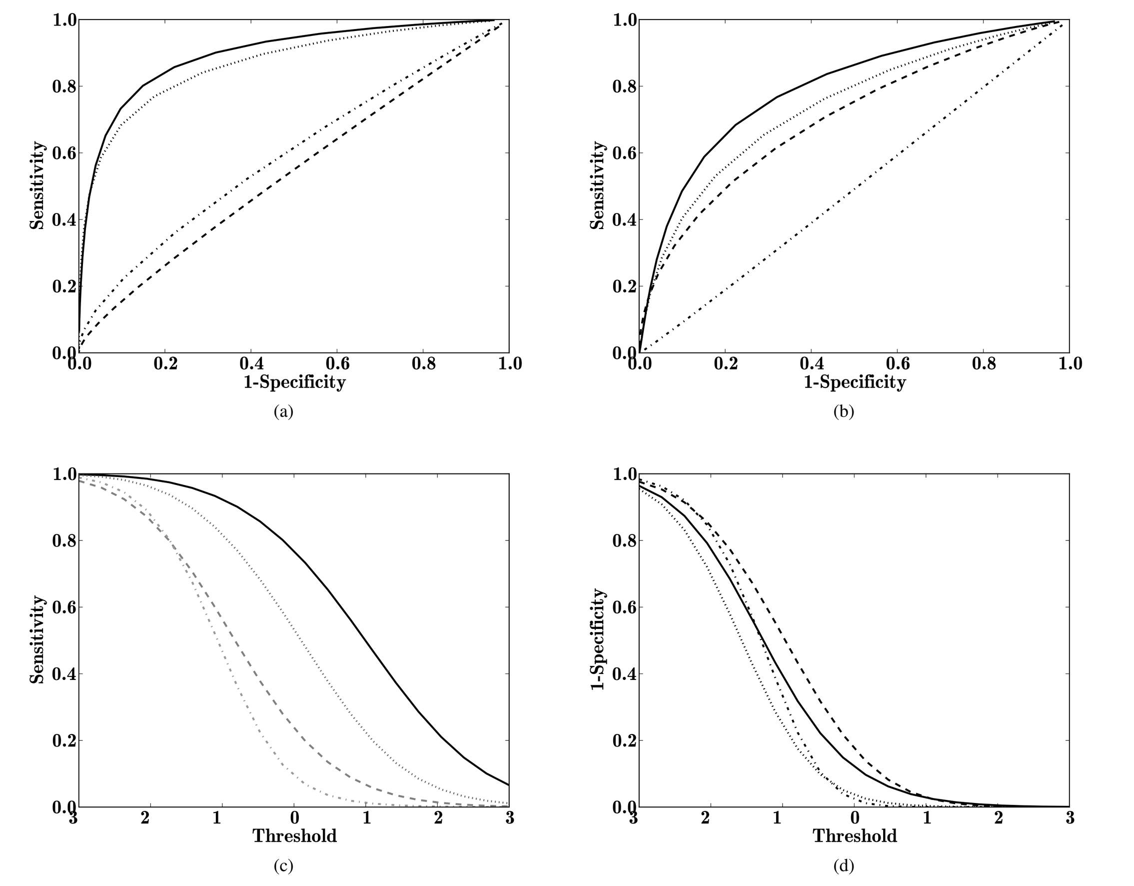


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

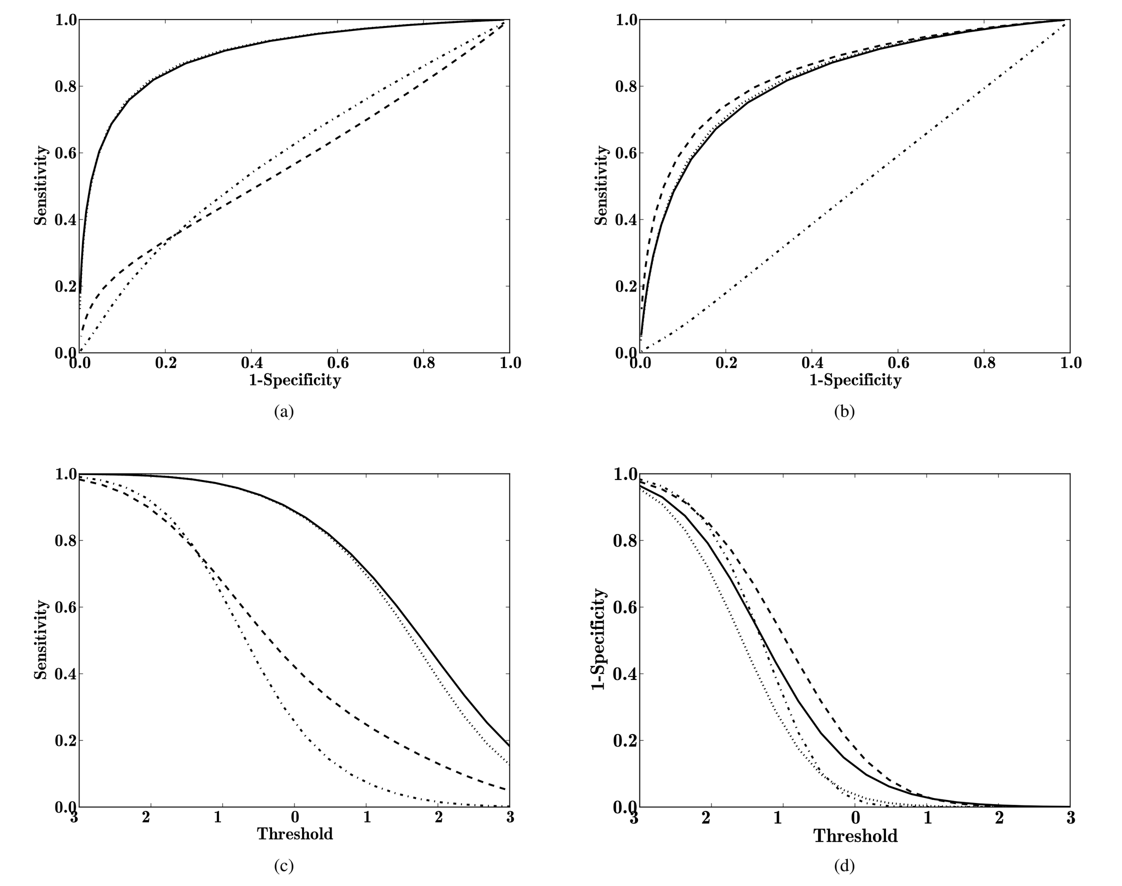


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text has become a more prevalent research topic only in the last few years. A subset of earlier work proved various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

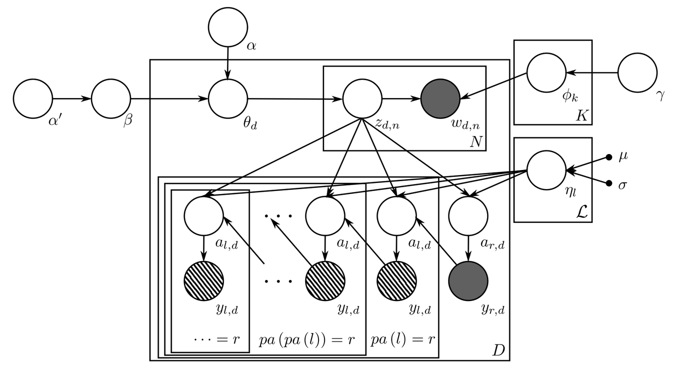


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

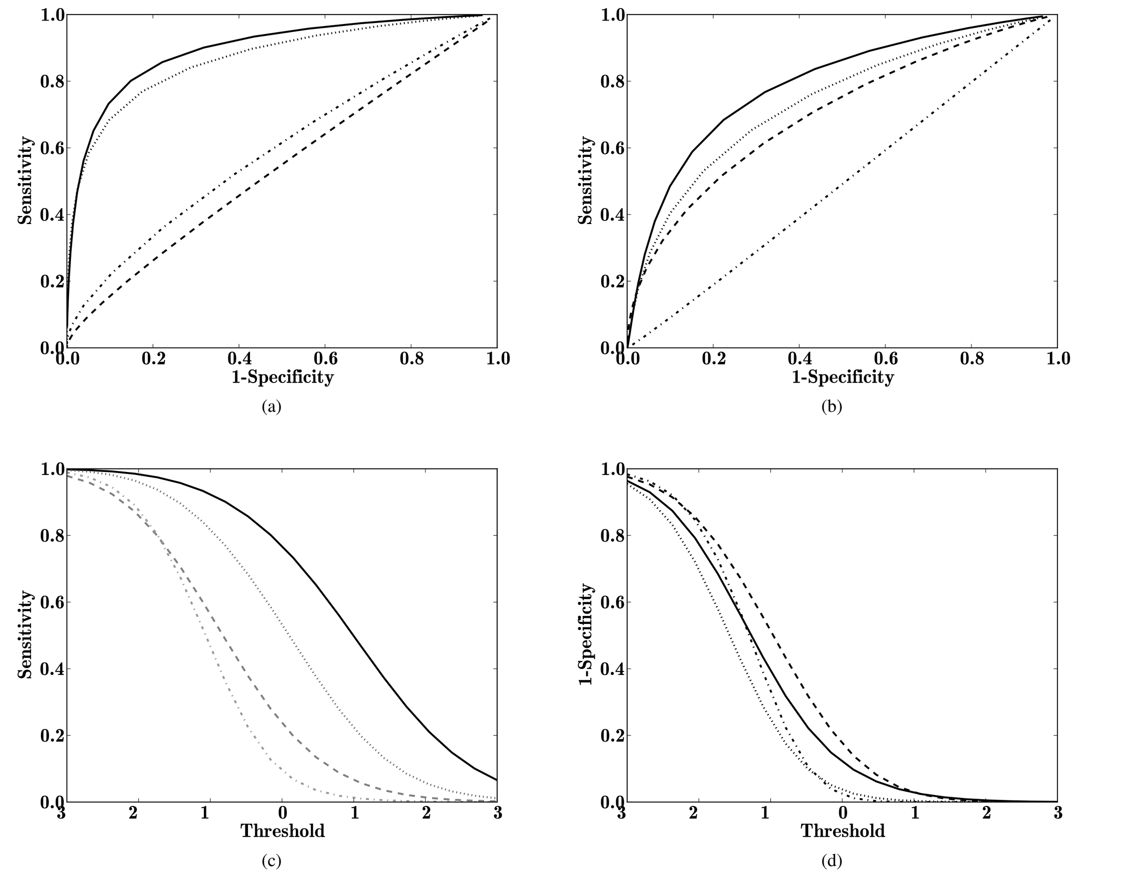


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

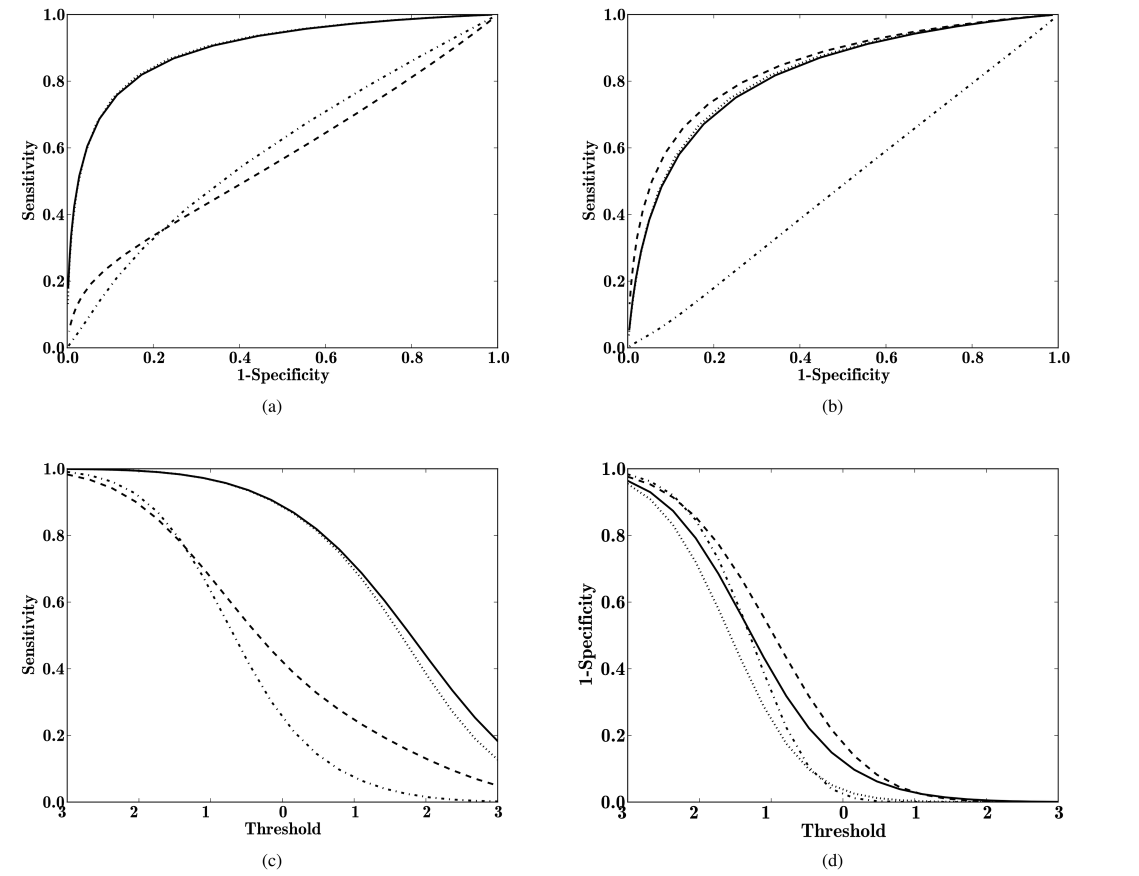


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggest that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

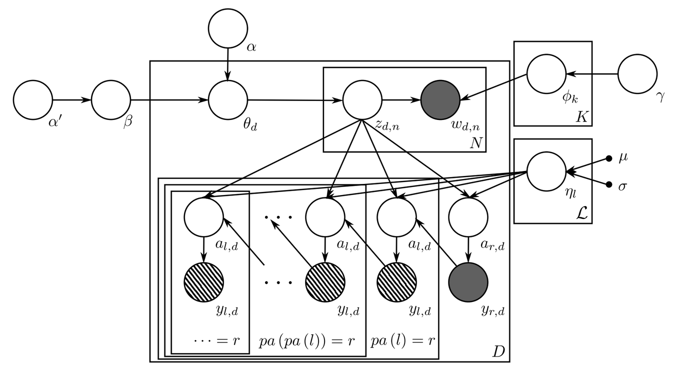


Figure 1: adapted sLDA model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{pa(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l), 1), & y_{pa(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l), 1)\mathbb{I}(a_{l,d} < 0), & y_{pa(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{pa(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior distribution.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \setminus z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{n,d} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

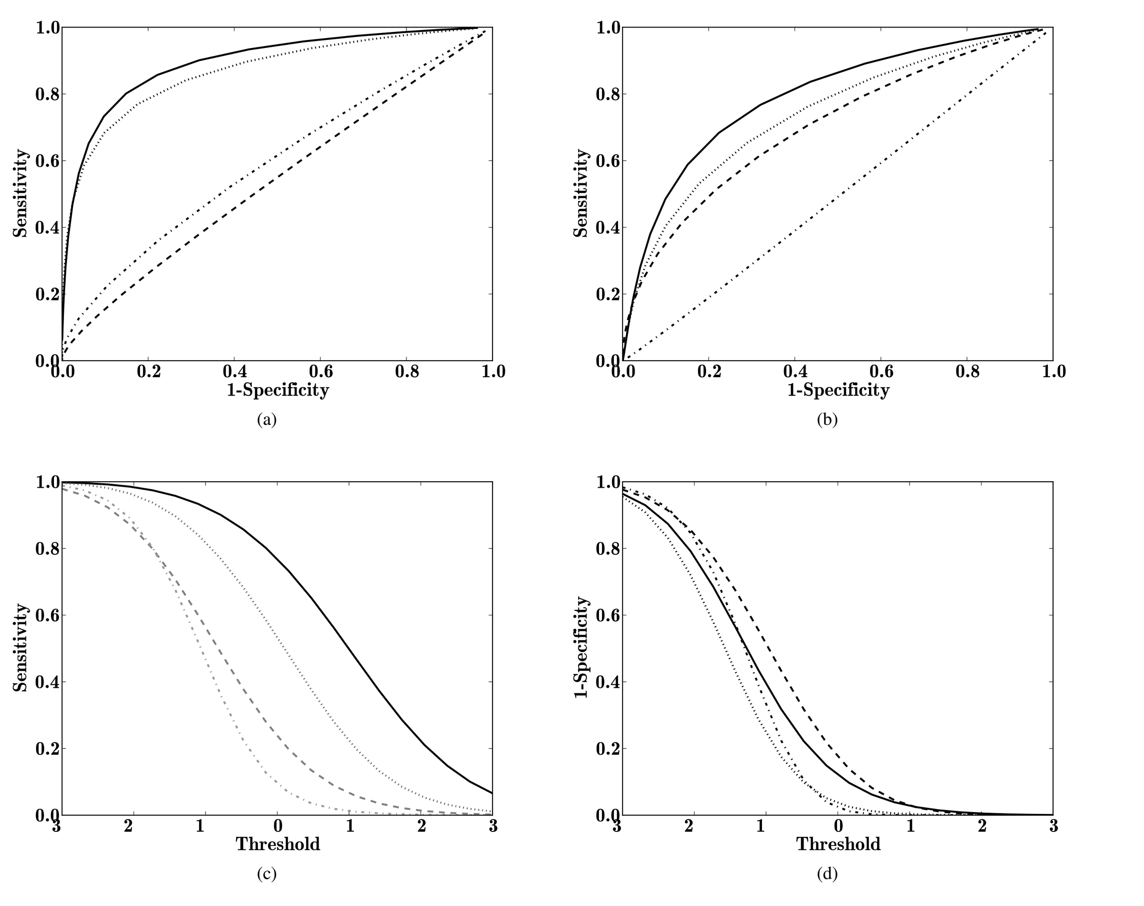


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

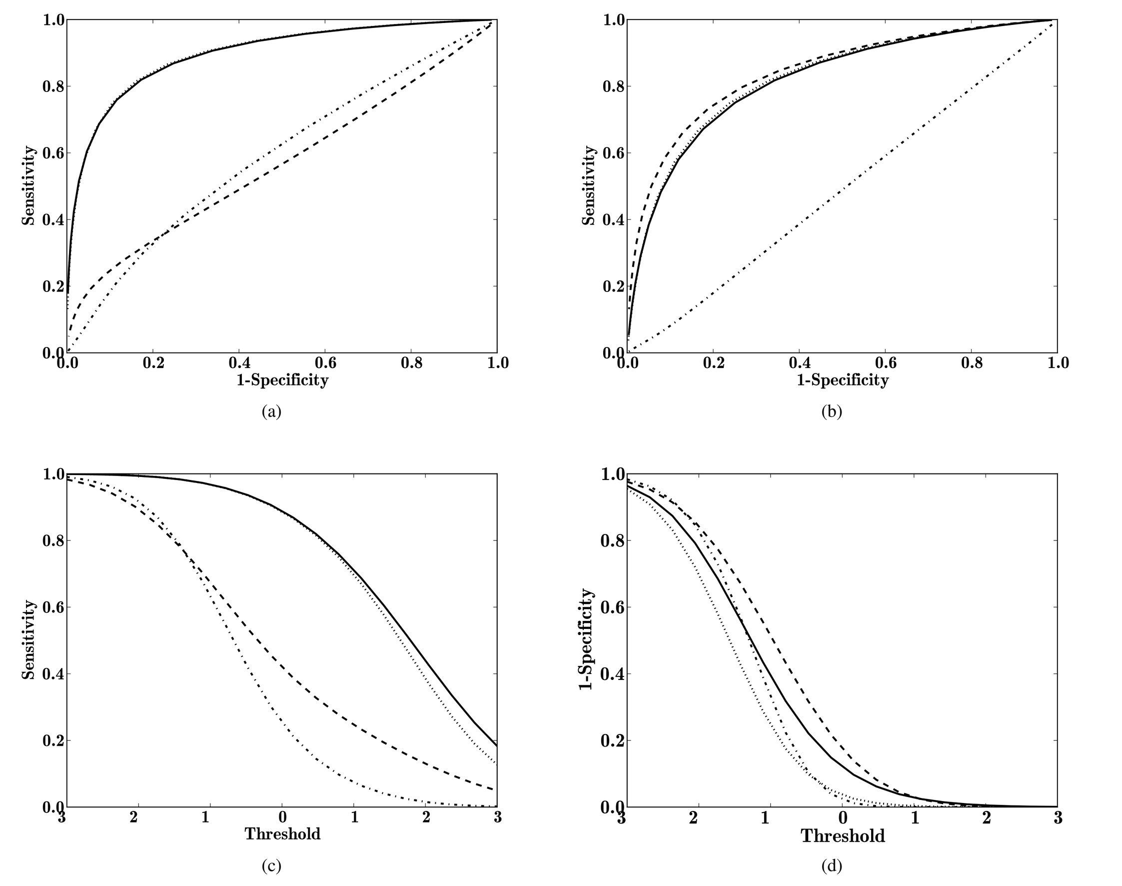


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[11] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[12] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[13] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 5 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

1.1 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [20, 12, 19, 7], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [18] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: "International Challenge: Classifying Clinical Free Text Using Natural Language Processing" (website). Most of the classification strategies included word matching and rule-based algorithms. [13, 9, 11]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [16]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

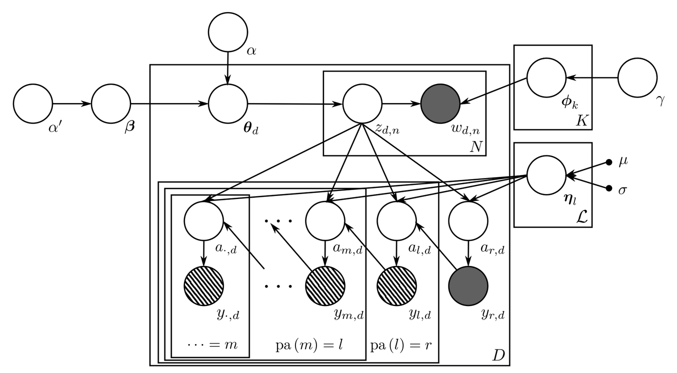


Figure 1: adapted sLDA model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:

     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\phi_{\mathbf{1:K}}$ and $\theta_{\mathbf{1:D}}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(n_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{n_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(n_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \quad (2)$$

Here, $n_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(\mathbf{d},\mathbf{n})}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3)$$
$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (6)$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \quad (7)$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right)$, $p\left(\alpha'\right)$, $p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

4.1 Data and Pre-Processing

4.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

4.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

4.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

4.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

5 Results

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

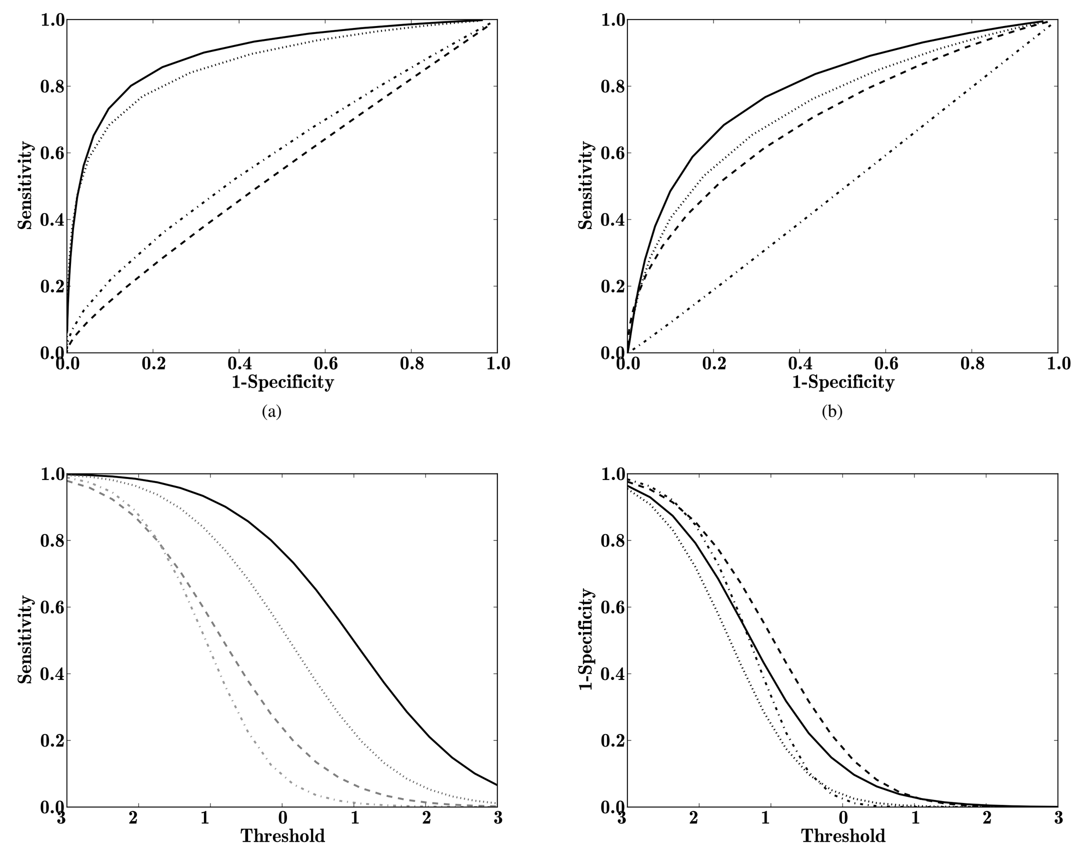


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

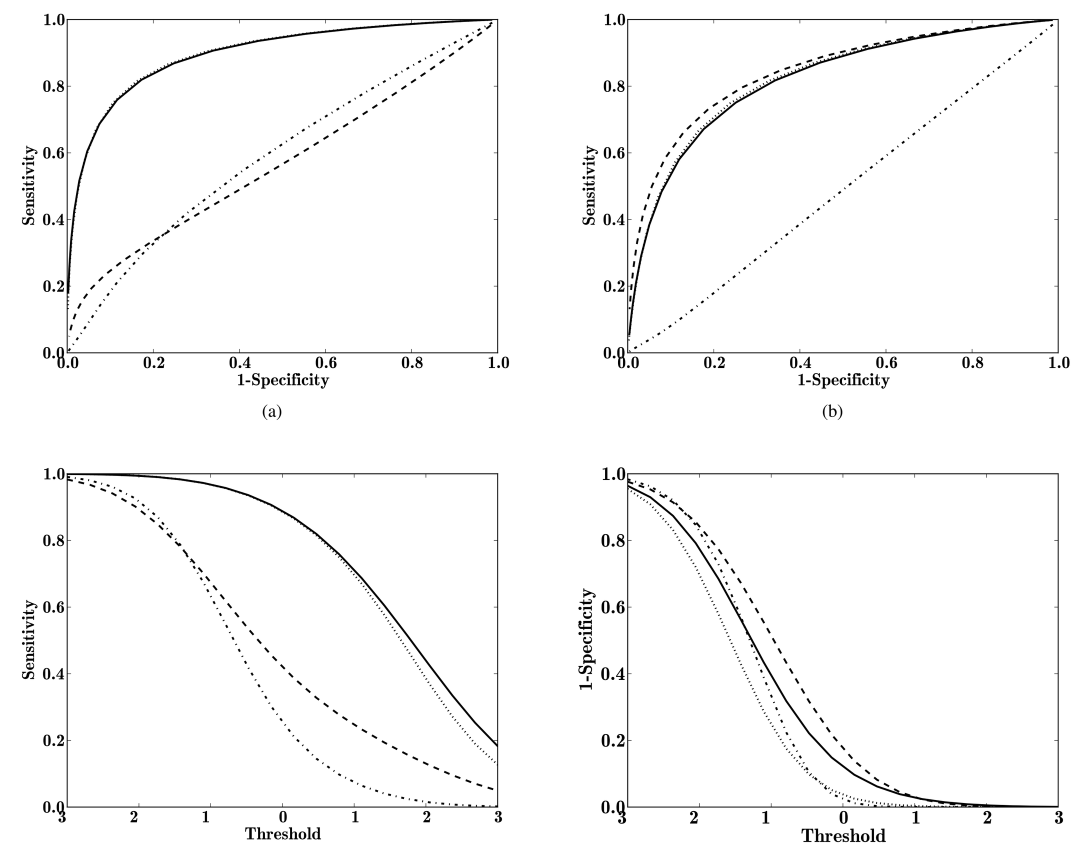


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggest that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 1.1 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 4 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

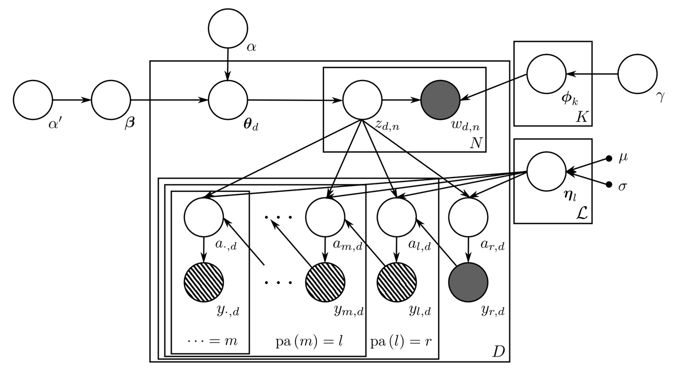


Figure 1: adapted sLDA model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

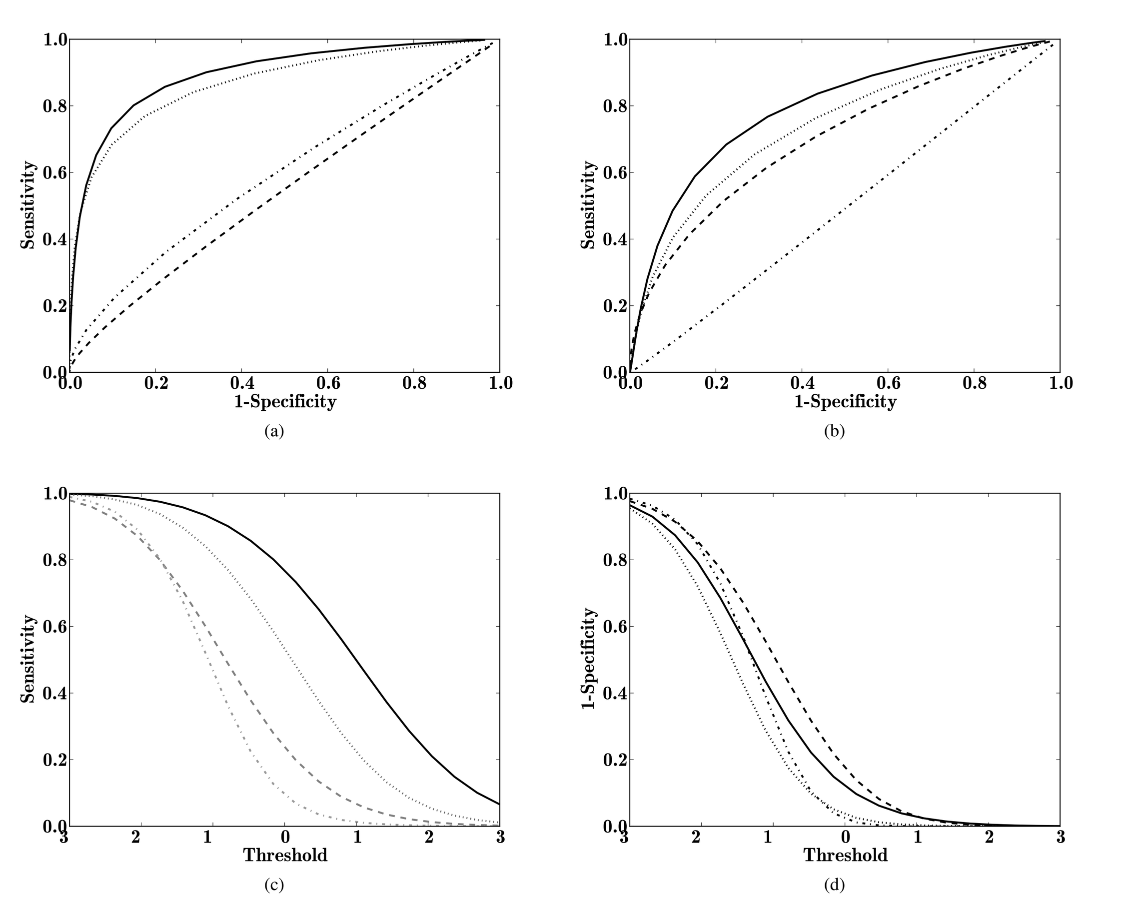
We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

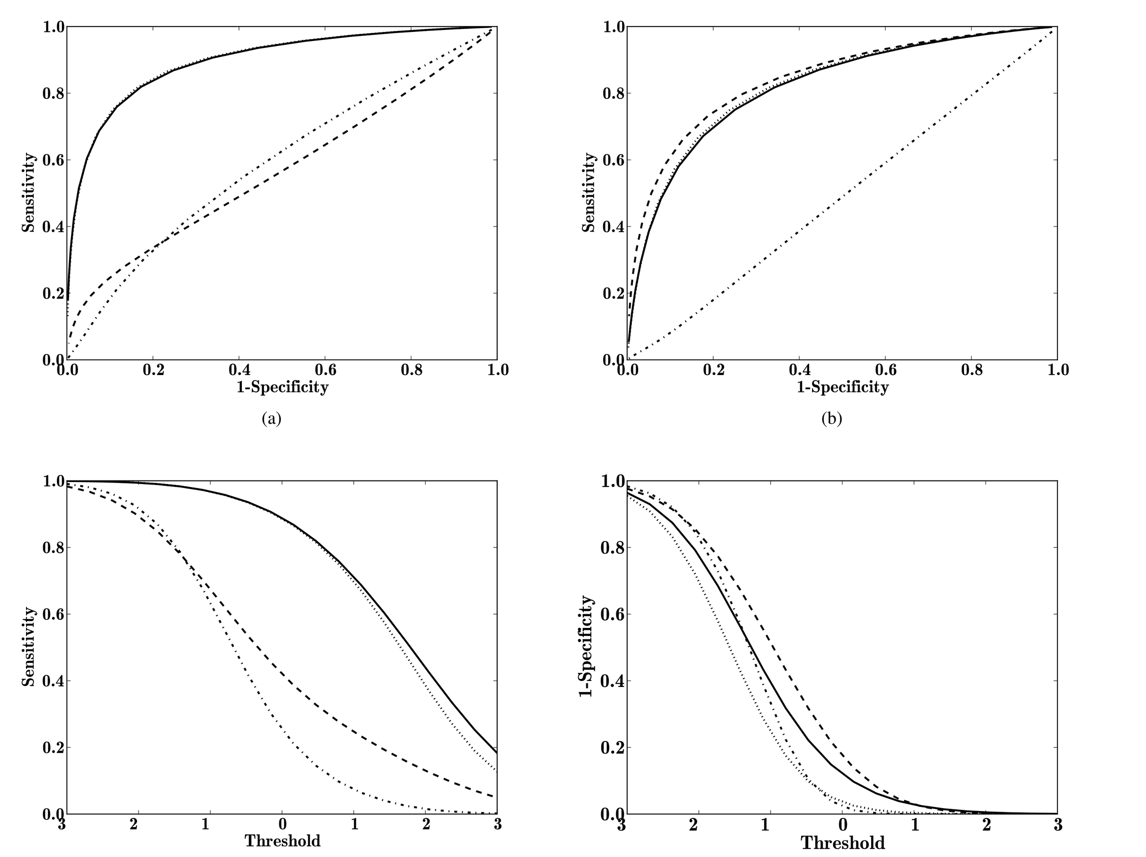


(a) (b) (c) (d)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.



(a) (b) (c) (d)

Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven crosslanguage retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
`{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu`

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured textual data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 5 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

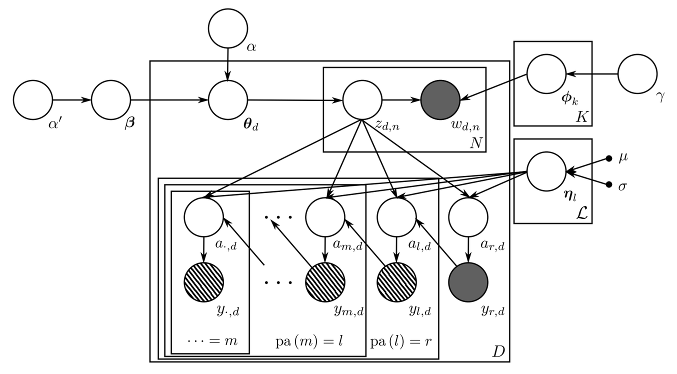


Figure 1: HSLDA graphical model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable
$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [14]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [14]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [14] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$
$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right)$, $p\left(\alpha'\right)$, $p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [17, 15, 23, 8]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [10]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

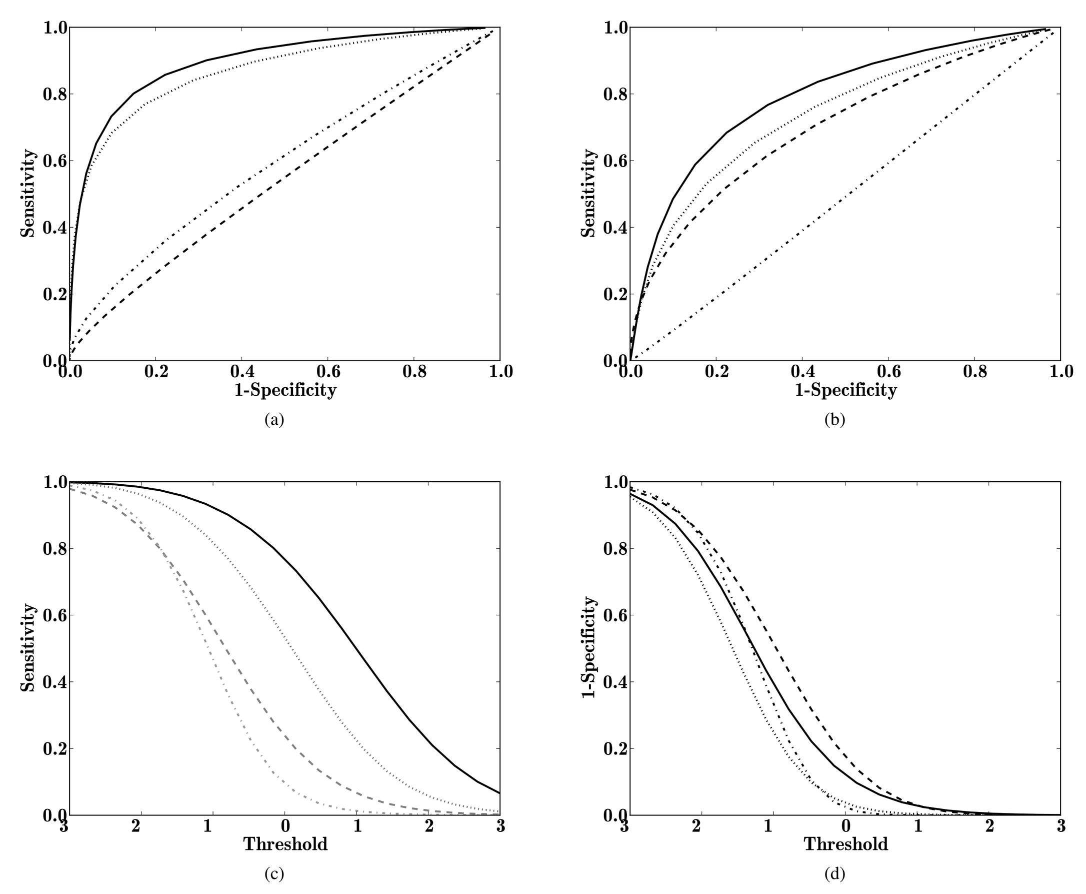


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

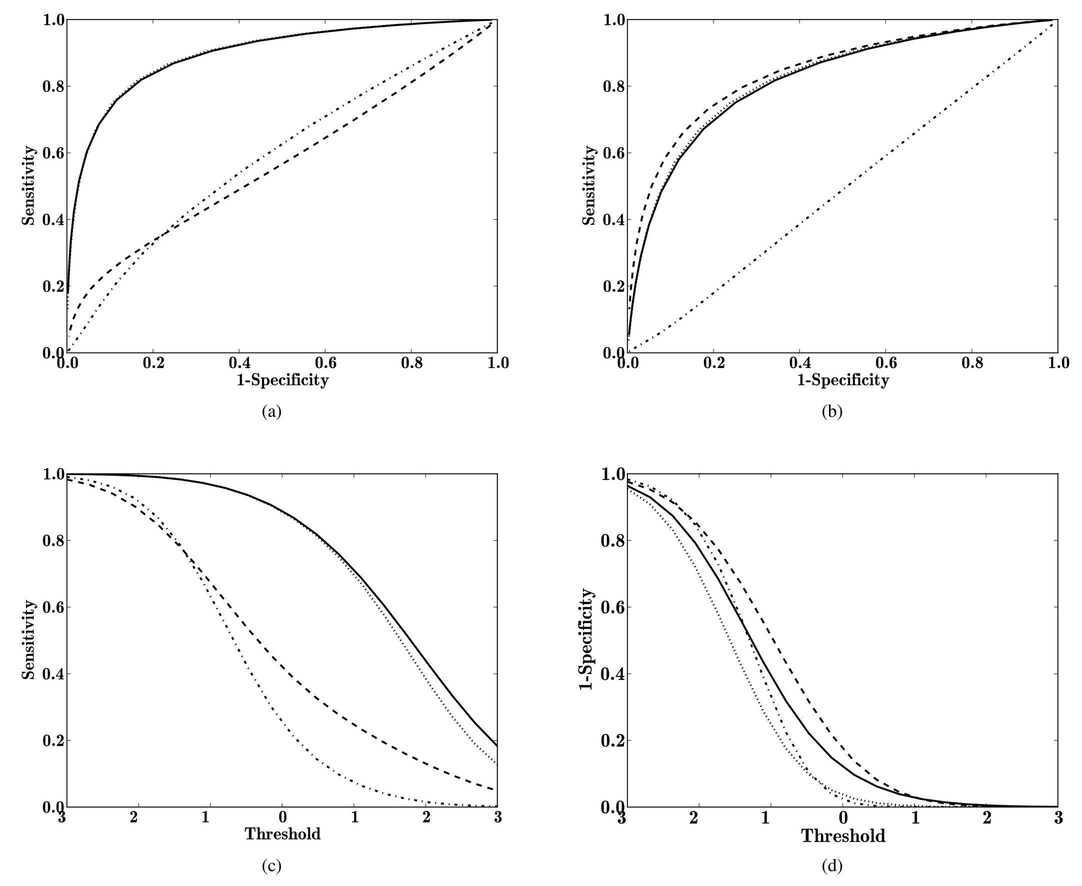


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. `http://www.amazon.com/`, 2011.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[15] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggest that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 5 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

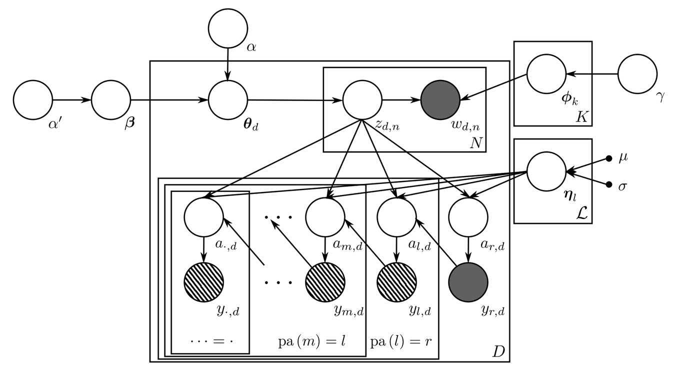


Figure 1: adapted sLDA model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because they are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ $\forall\, l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

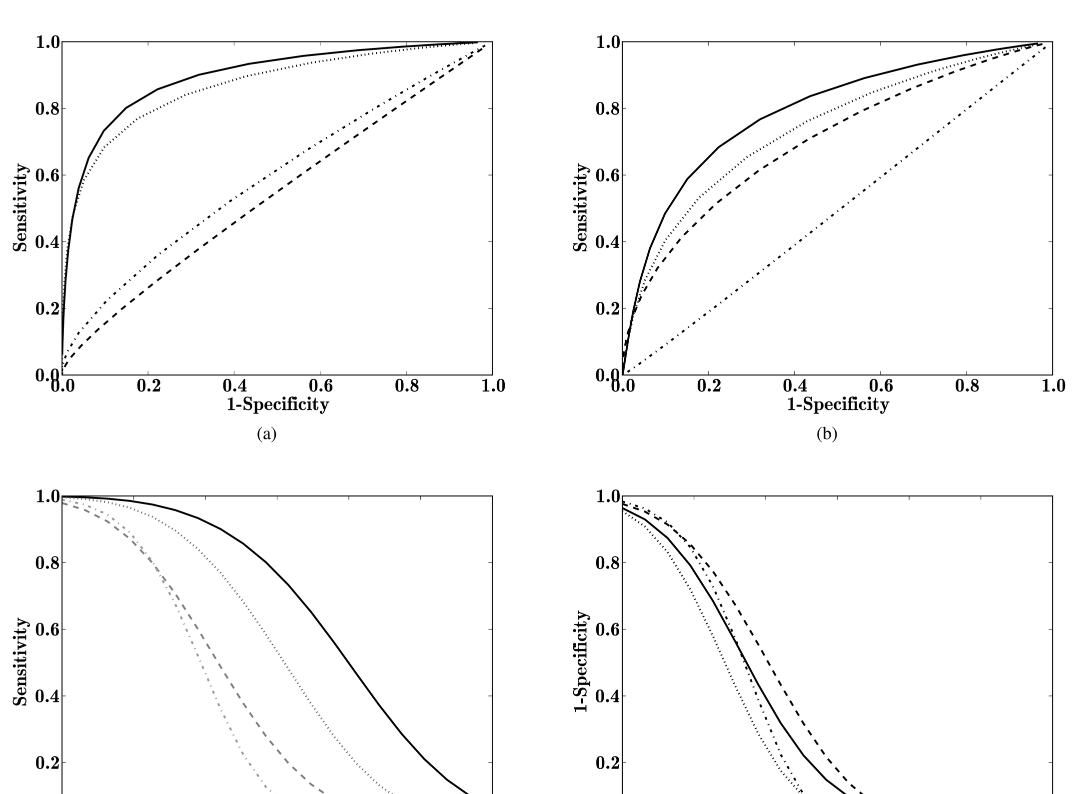


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

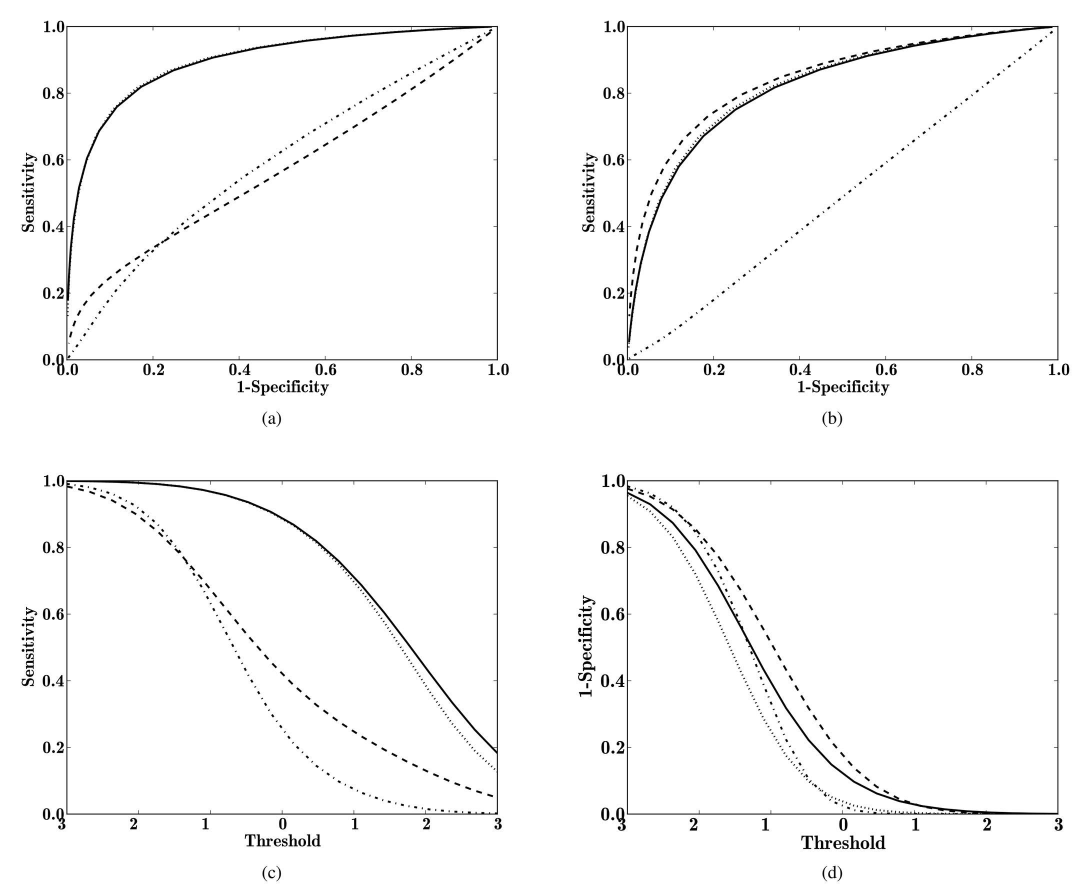


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 5 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hieararchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

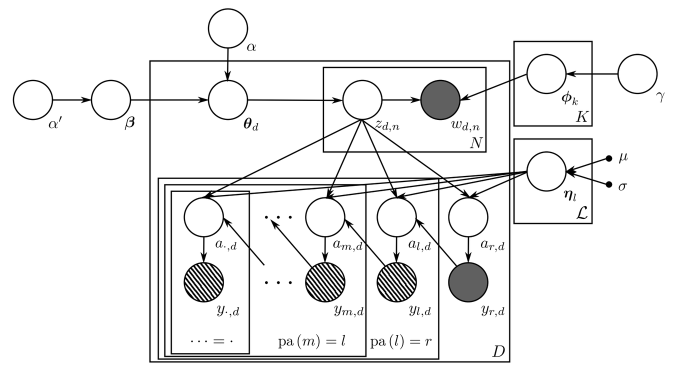


Figure 1: HSLDA graphical model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \prod_{l \in \mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l)\, p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\}. \quad (2)$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3)$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (6)$$

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma(\alpha \boldsymbol{\beta}_k + n_{d,k})} s(n_{d,k}, m) (\alpha \boldsymbol{\beta}_k)^m \quad (7)$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

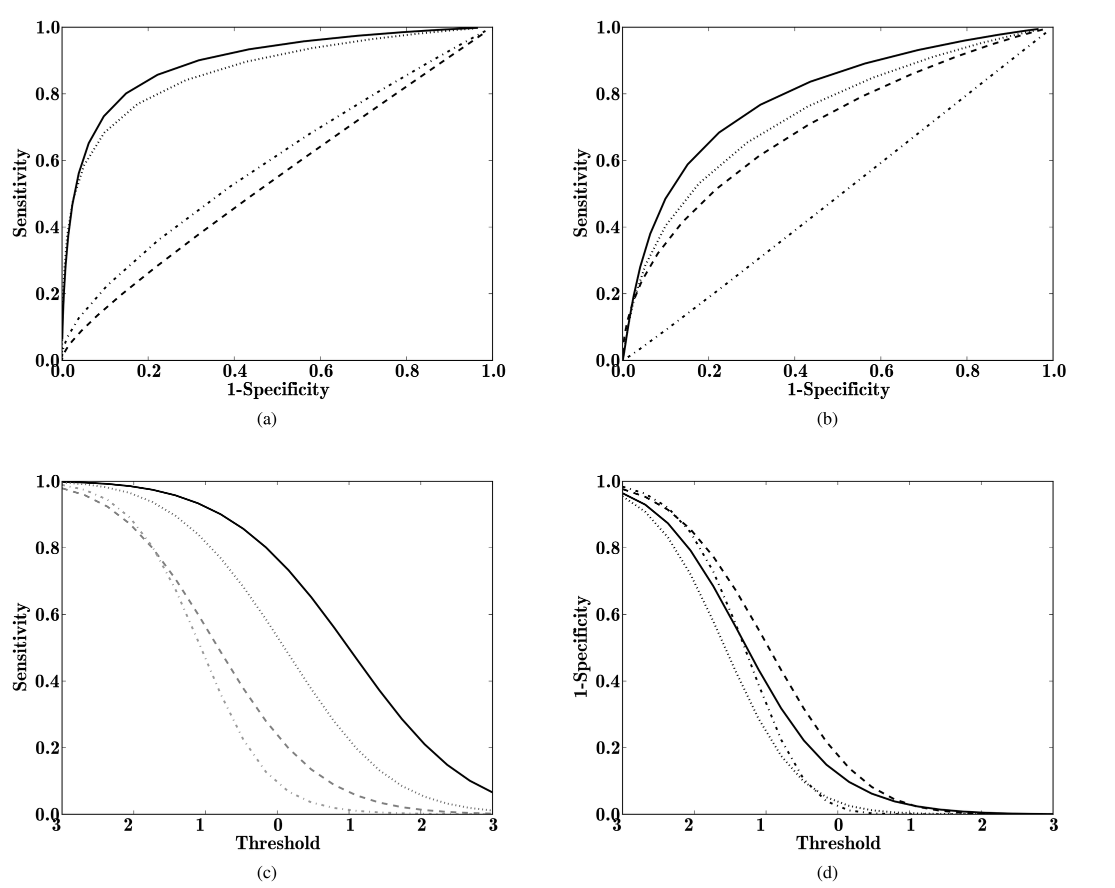


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

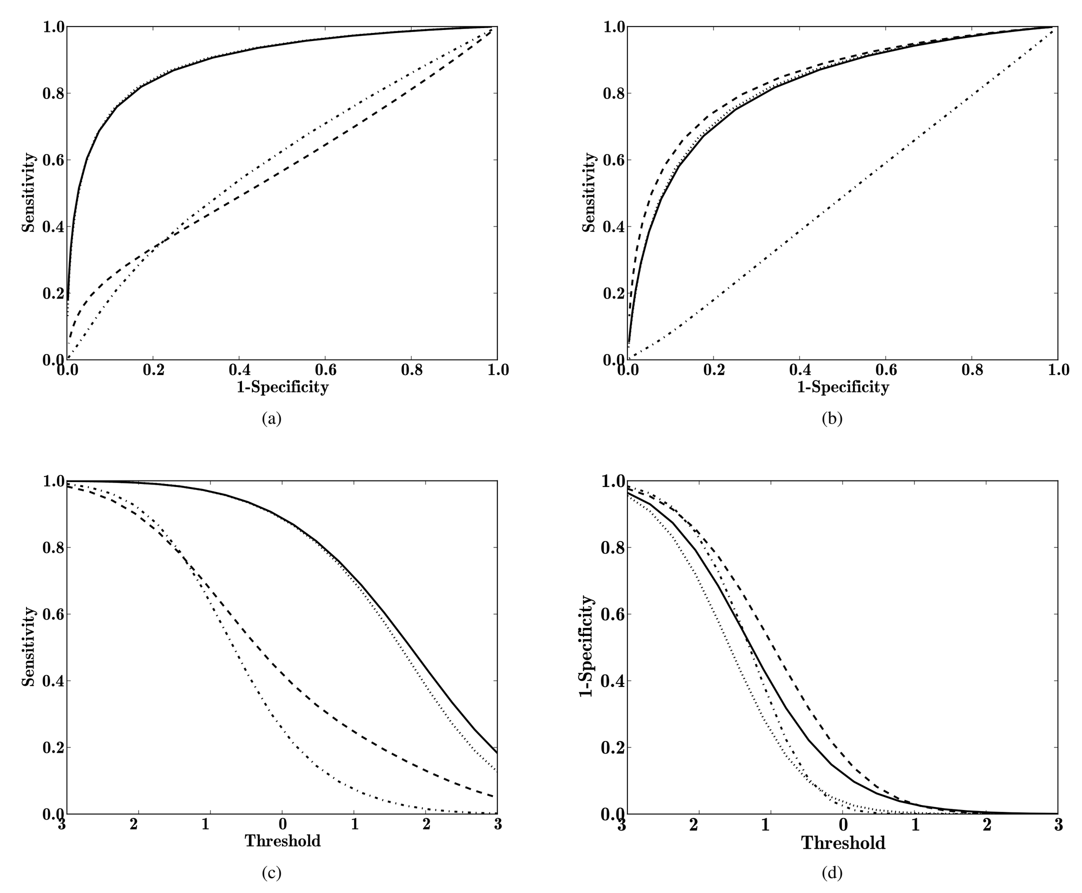


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.
[2] DMOZ open directory project. http://www.dmoz.org/, 2002.
[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.
[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.
[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.
[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.
[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.
[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.
[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.
[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.
[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision", often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to health record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply labeled, bag-of-words data. We will refer to the grouped bag-of-word data as a document. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the ordered set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

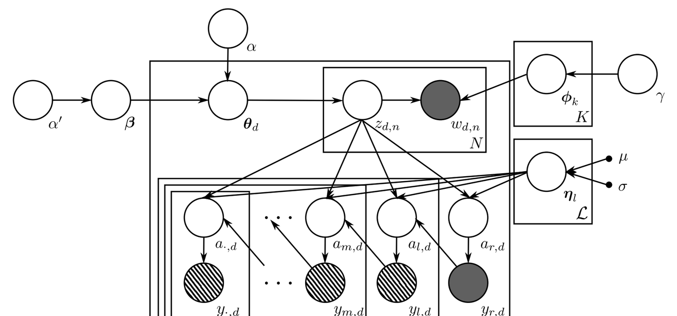


Figure 1: HSLDA graphical model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\beta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable

normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\mu}$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\mu} = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models

as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

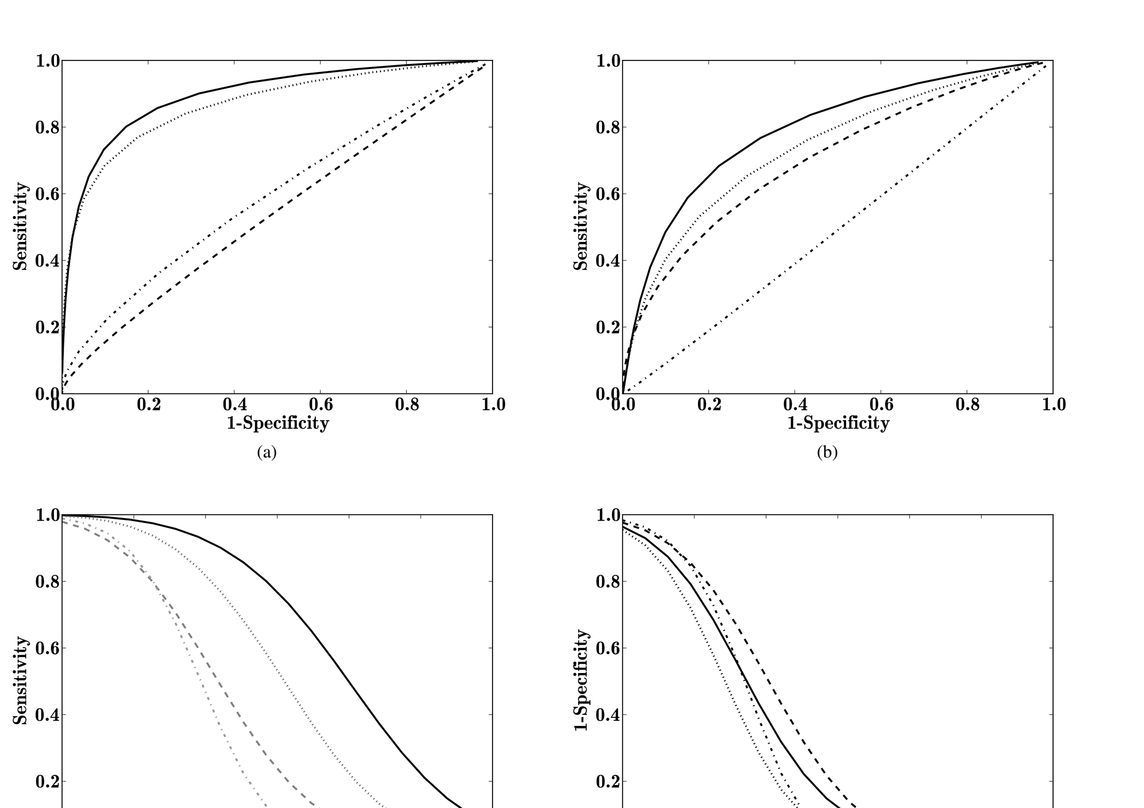


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

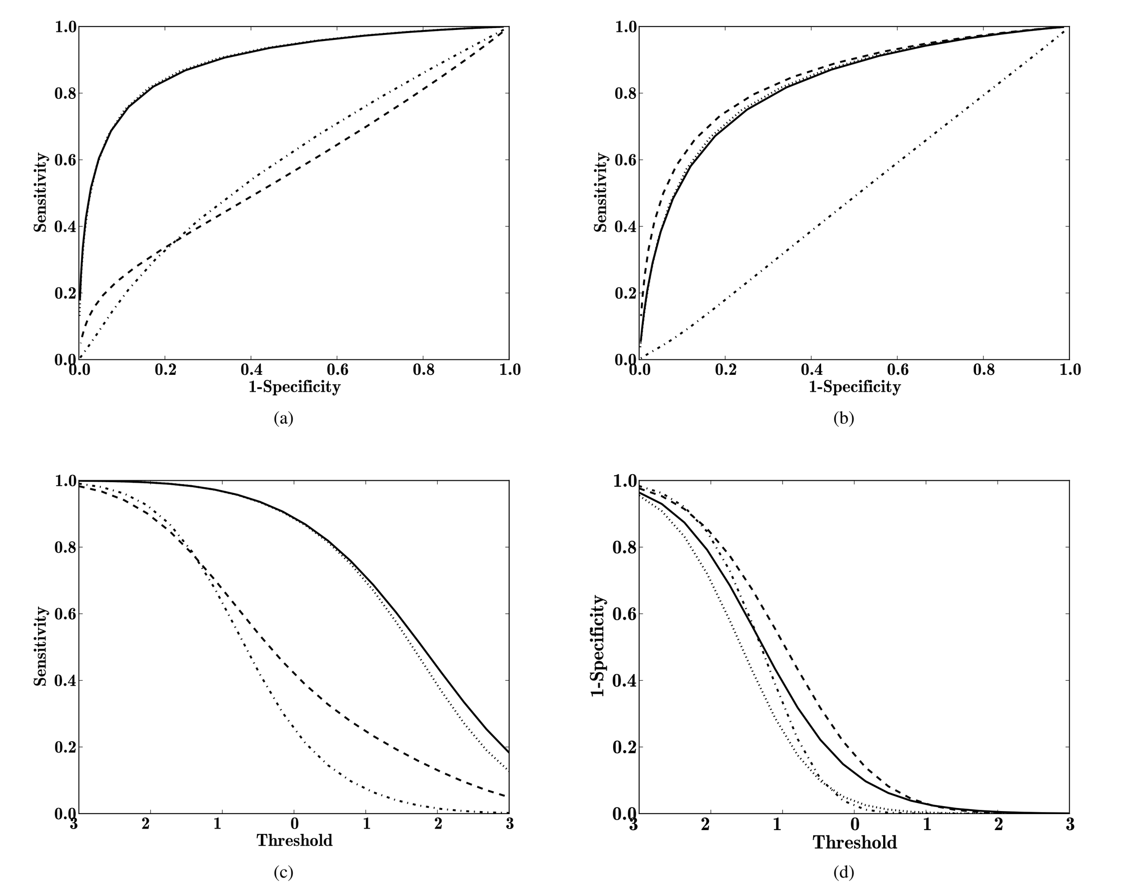


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-words data. Out-of-sample label prediction is the primary goal of this work; however, improved dimensionality reduction is also of interest. We define a model that uses probit regressors on a conditionally dependent label hierarchy tied to latent Dirichlet allocation (LDA). We find that the additional signal that comes from multiple, hierarchically constrained labels substantially improves out-of-sample label prediction in comparison to supervised LDA approaches that don't utilize information derived from the structure of the label space. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topic model also improves as a result of using this signal.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this "supervision" can be seen as extra data about a document; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally draw given an inferred document-specific topic mixture. It has been demonstrated that the signal provided by this supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesis that hierarchical labels should, at least in theory, provide better signal than the simpler unstructured supervision previously considered. Results from applying our model to medical record and web retail data suggests that this is likely to be the case. In particular, we observed big gains in our primary goal of out-of-sample label prediction when using hierarchical supervision.

The remainder of this paper is structured as follows. In Section 4 we review related work, in Section 2 we introduce hierarchically supervised LDA (HSLDA), and in Section 5 we apply HSLDA to health care and web retail data, showing predictive performance and improved topic generation.

2 Model

We define here a hierarchically supervised LDA model. Although we will focus on document modeling in our description and experiments, this model applies equally well to other collections of discrete data with hierarchically constrained labels.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hieararchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

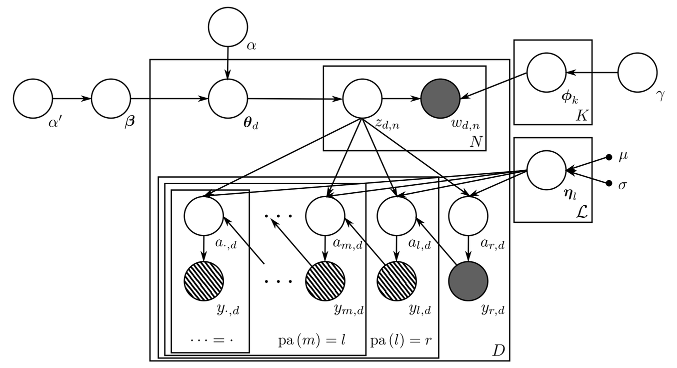


Figure 1: adapted sLDA model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \prod_{l \in \mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l)\, p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}^T_d \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\boldsymbol{\Sigma}$ such that

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$
$$\hat{\boldsymbol{\mu}}_l = \boldsymbol{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$
$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m) (\alpha\beta_k)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical information, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,d} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

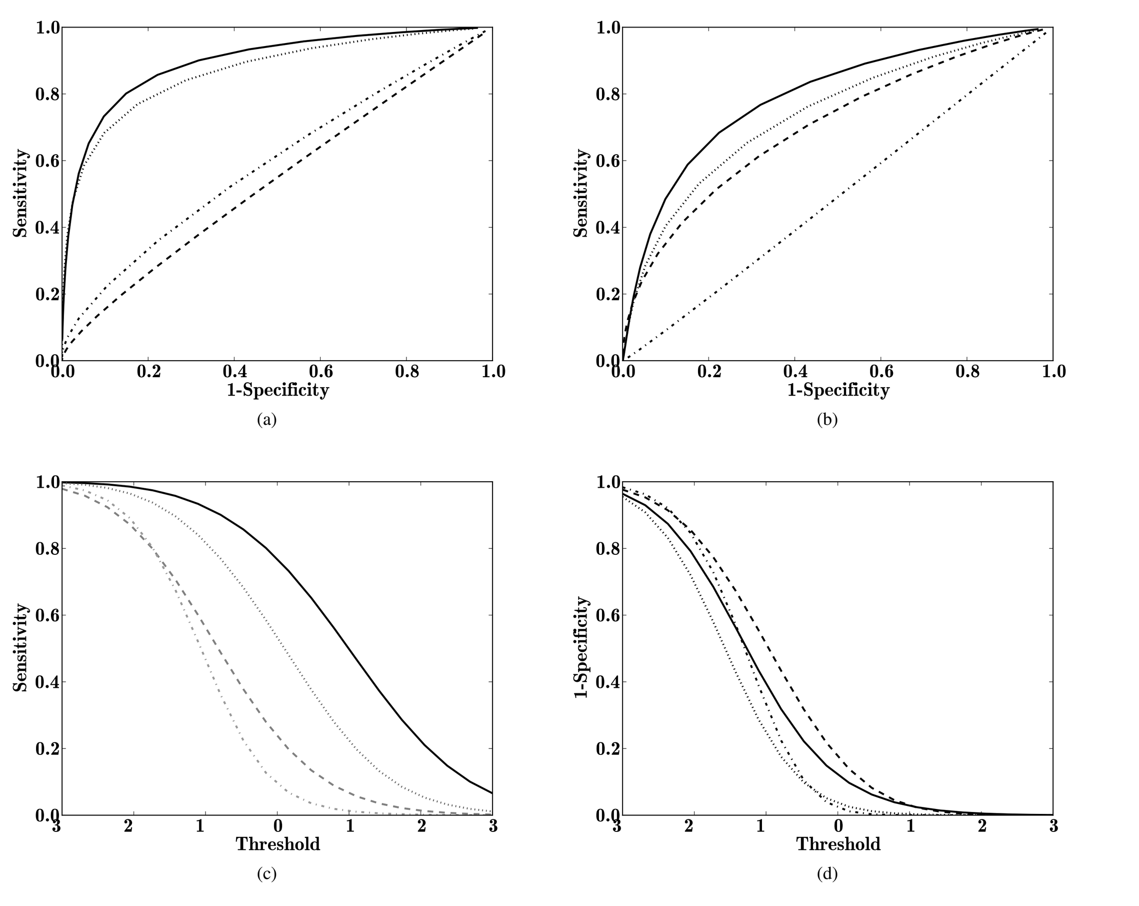


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

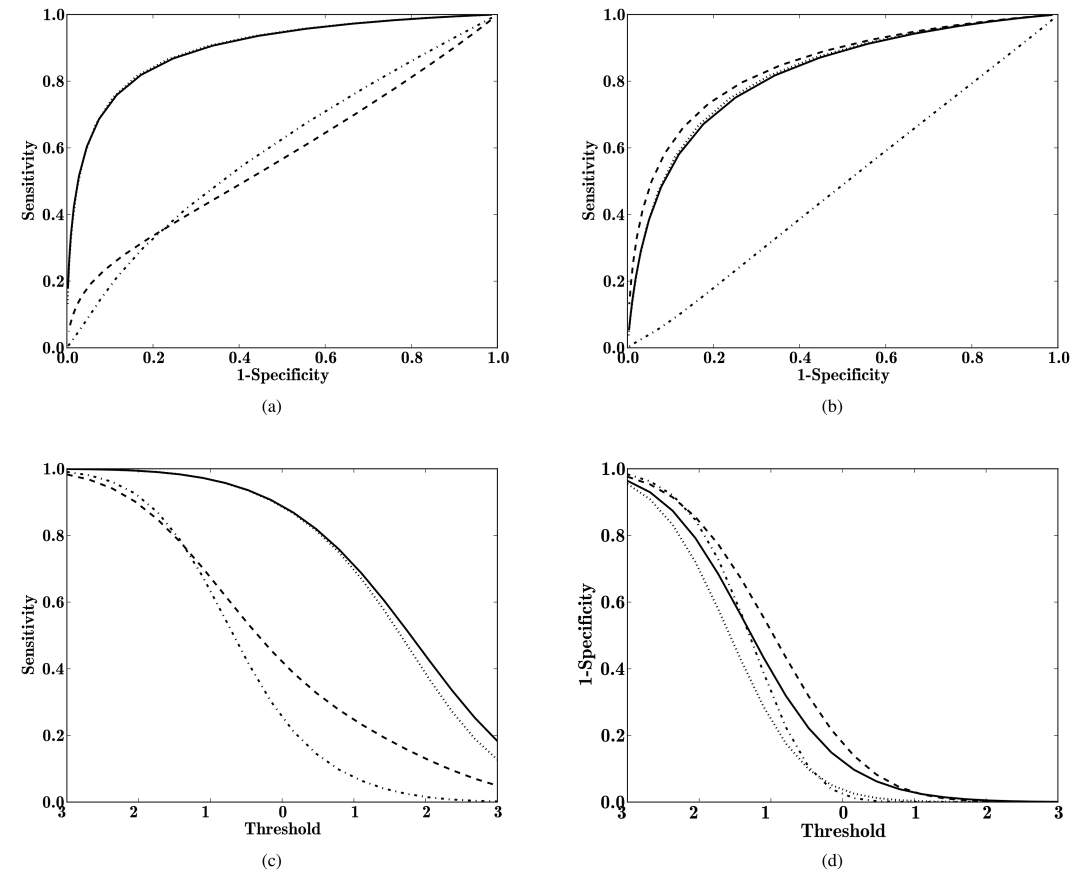


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply labeled, bag-of-words data. We will refer to the grouped bag-of-word data as a document. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the ordered set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$.

We assume a pre-specified set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l. The label set is assumed to be structured as an "is-a" hierarchy. To understand this, consider a hierarchy where label l_1 is a parent of label l_2. If document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

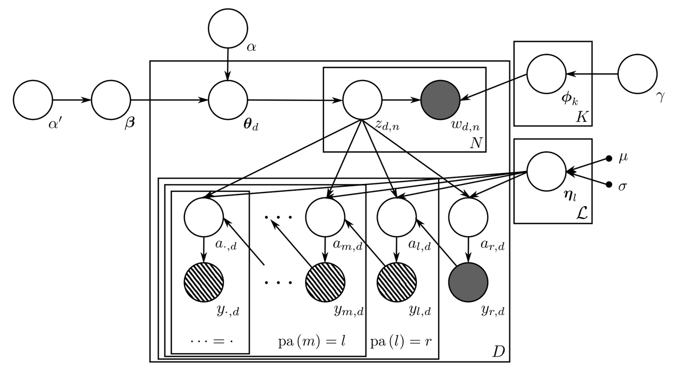


Figure 1: HSLDA graphical model

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

       where $\bar{z}_d = N_d^{-1}\sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L}\backslash\mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l\in\mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l'\in\mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable

normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l\in\mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l\in\mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T\boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T\bar{\mathbf{z}}_d\right)\right\}\mathbb{I}\left(a_{l,d}y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models

as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

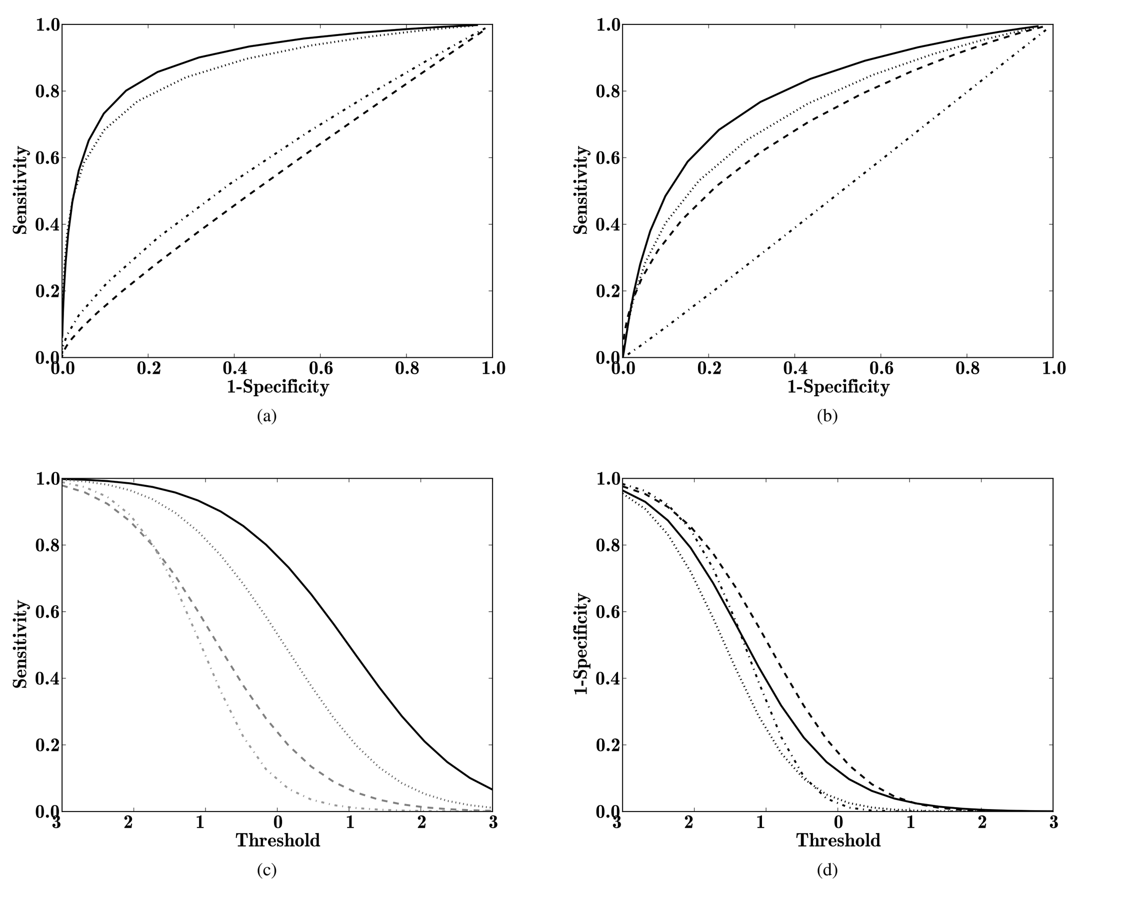


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

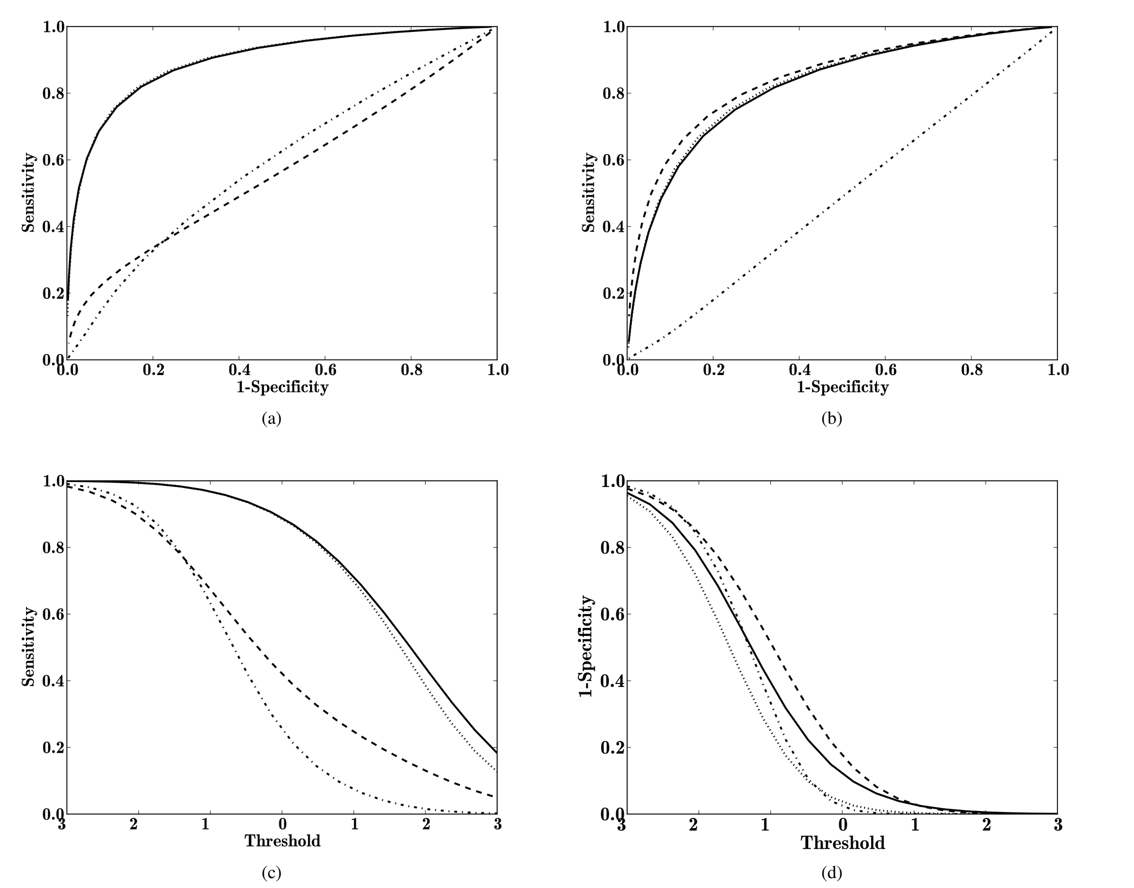


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \dots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

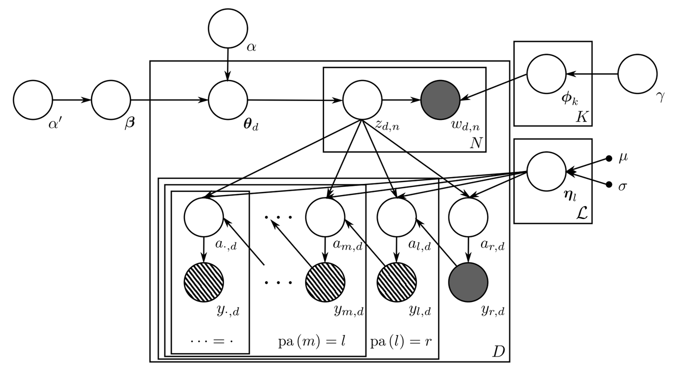


Figure 1: HSLDA graphical model

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \dots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \dots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \dots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing.

In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\dots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\dots,D}$, $\boldsymbol{\beta}$, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \dots, N_d$ in documents $d = 1, \dots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \prod_{l \in \mathcal{L}_d} p(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l)\, p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \dots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \dots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \dots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s(n_{d,k}, m) (\alpha\beta_k)^m \tag{7}$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \dots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

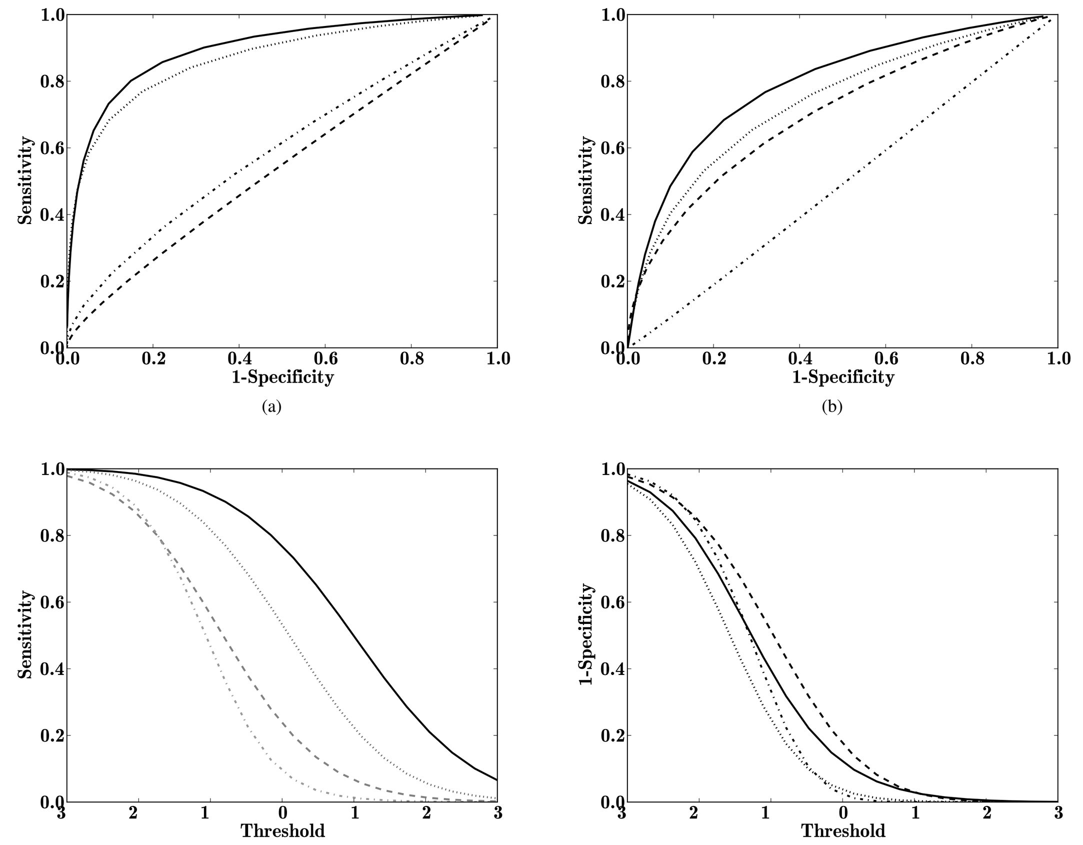


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

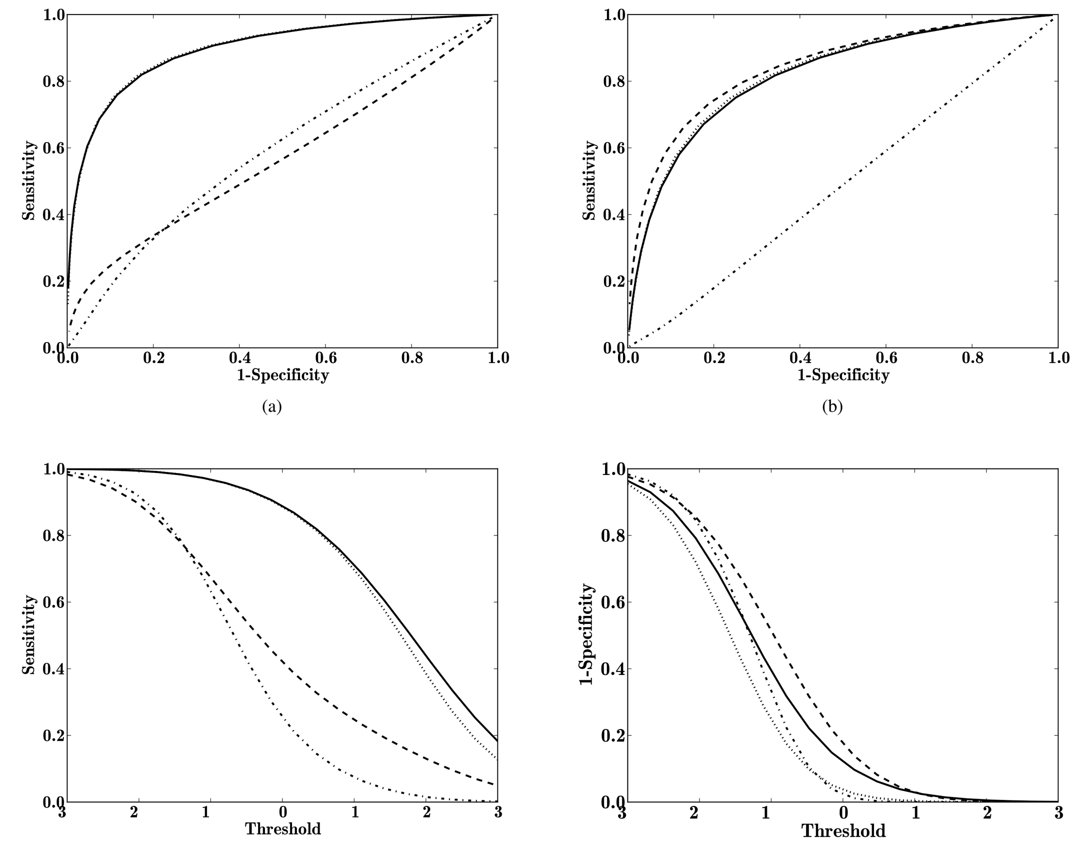


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

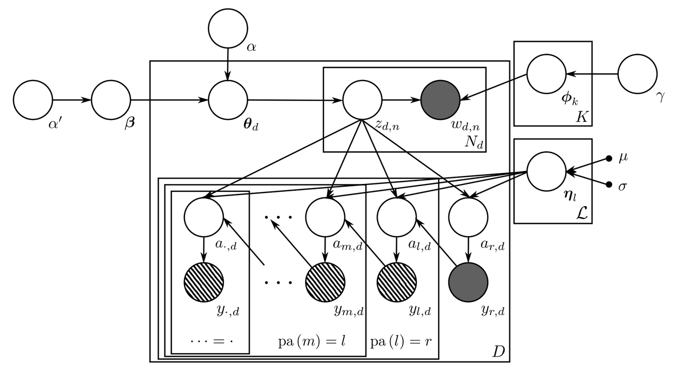


Figure 1: HSLDA graphical model

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     – Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     – Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     – Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     – Set the response variable
     $$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing.

In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$
$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.
[2] DMOZ open directory project. http://www.dmoz.org/, 2002.
[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.
[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.
[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.
[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.
[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.
[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.
[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.
[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.
[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.
[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.
[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

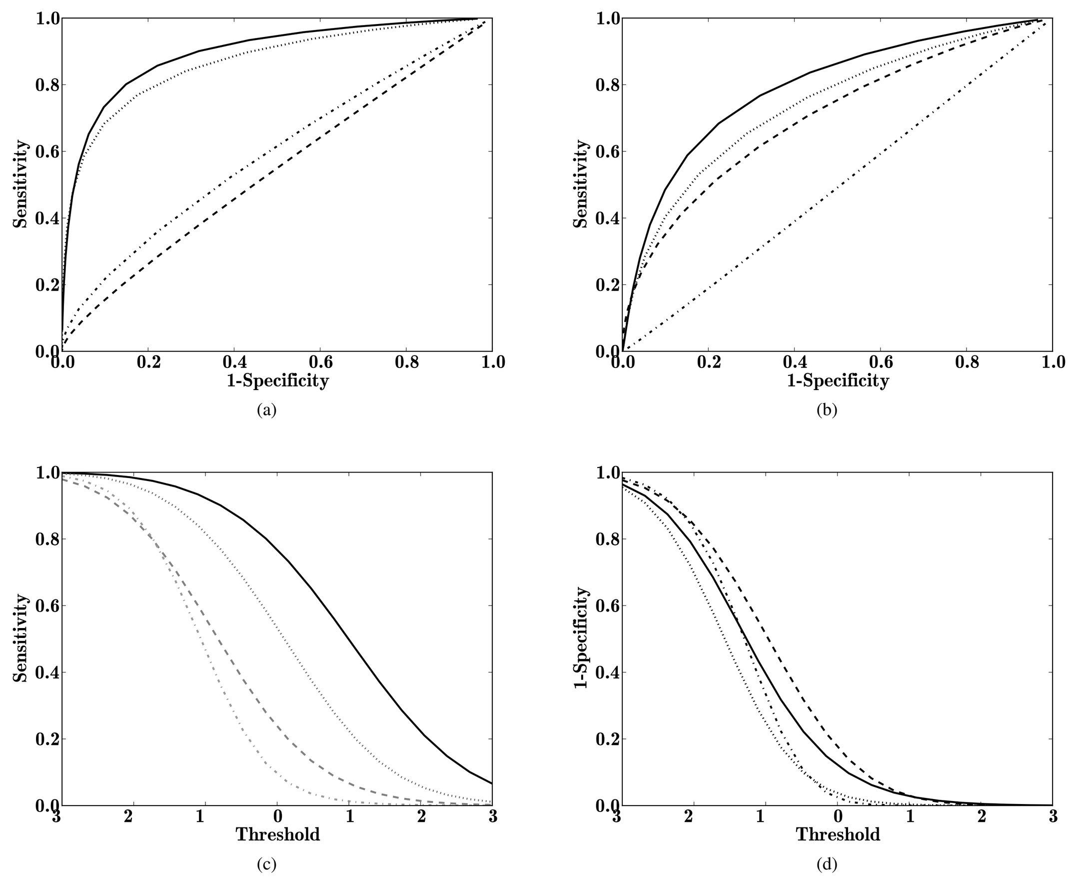


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

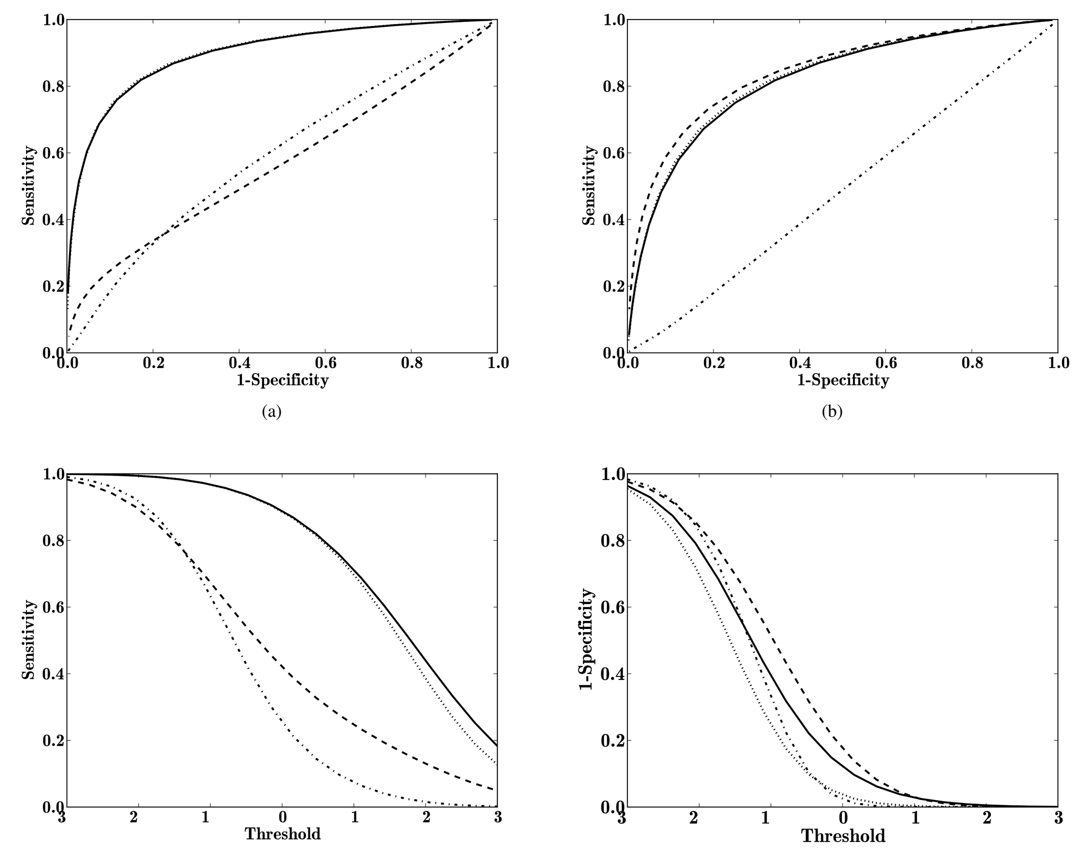


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision", often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

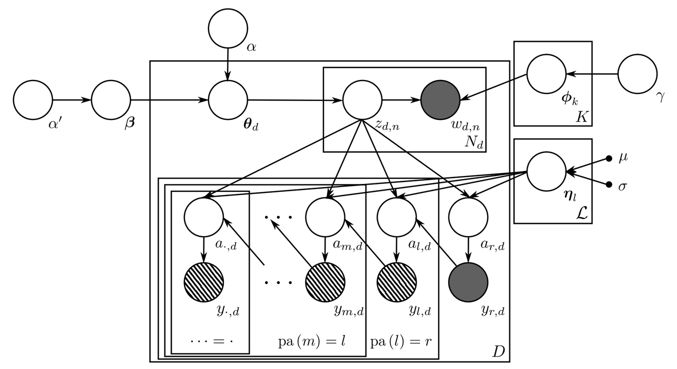


Figure 1: HSLDA graphical model

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing.

In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}$, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \setminus z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,d} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] ...

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

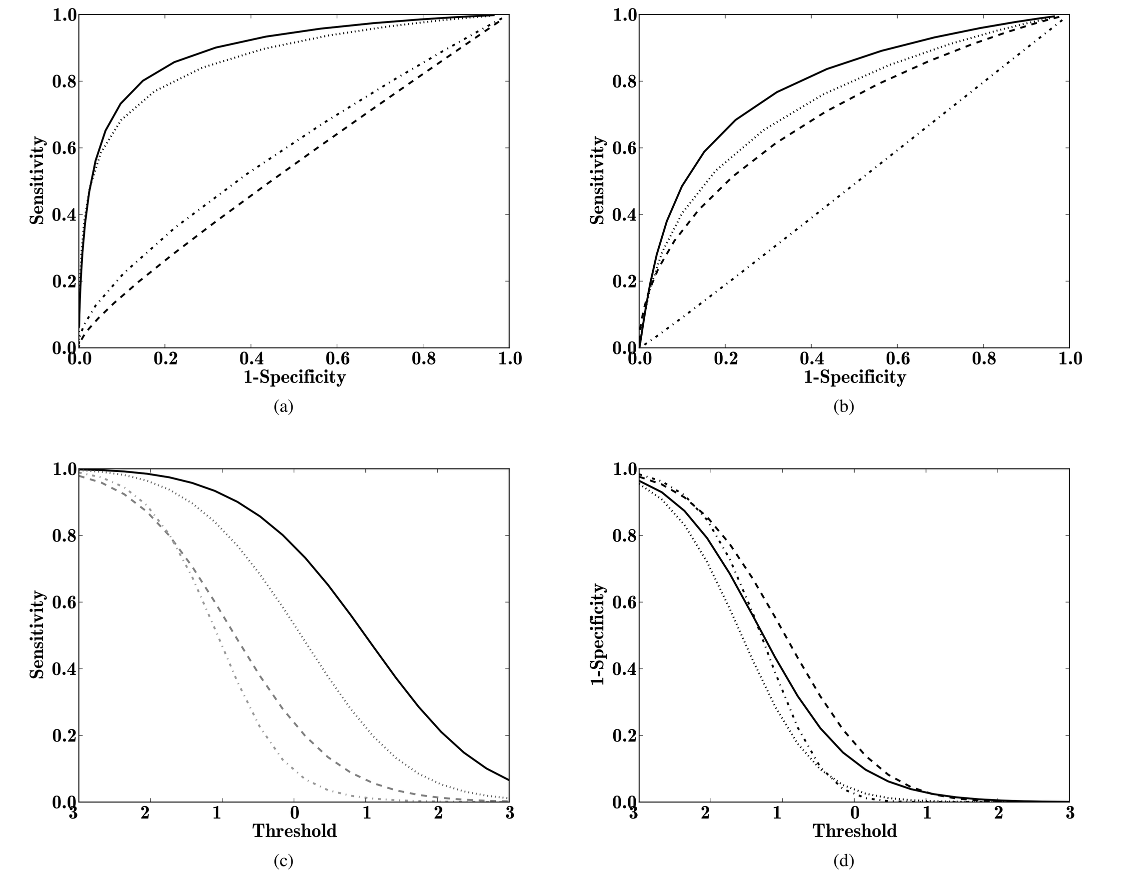


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

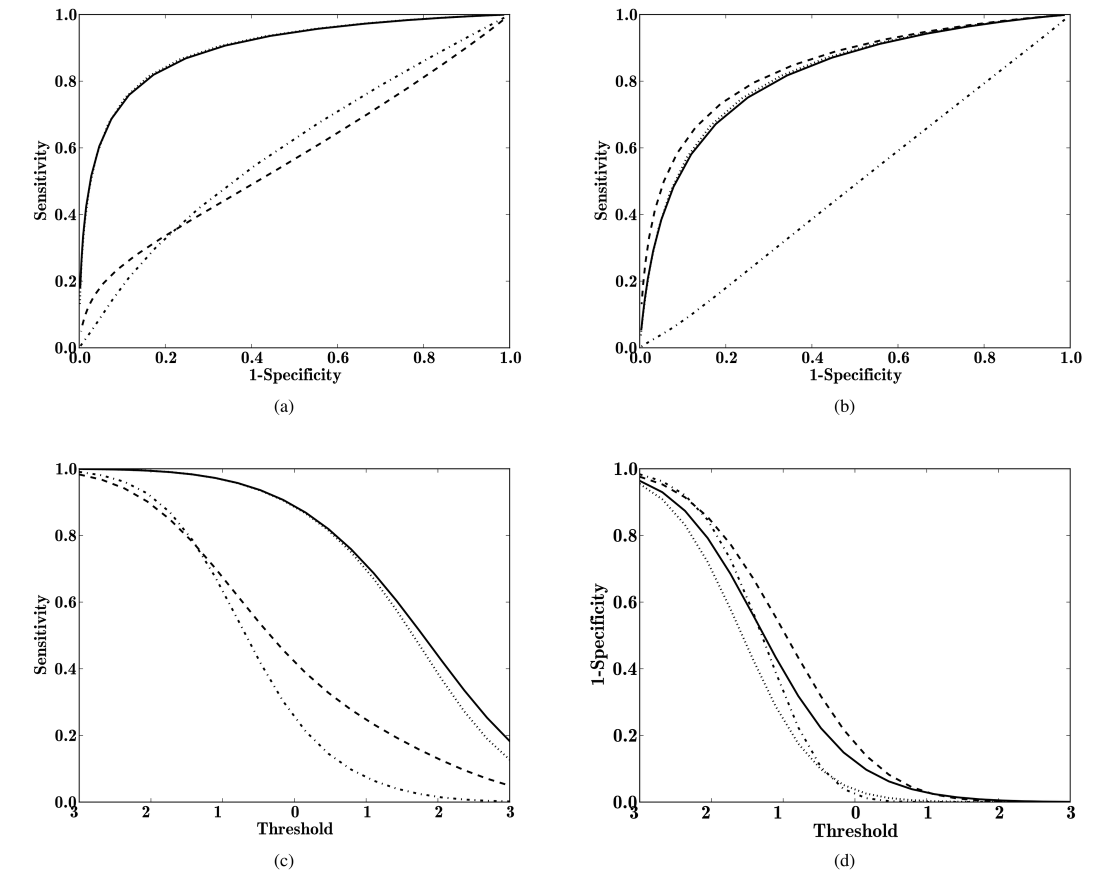


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noemie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

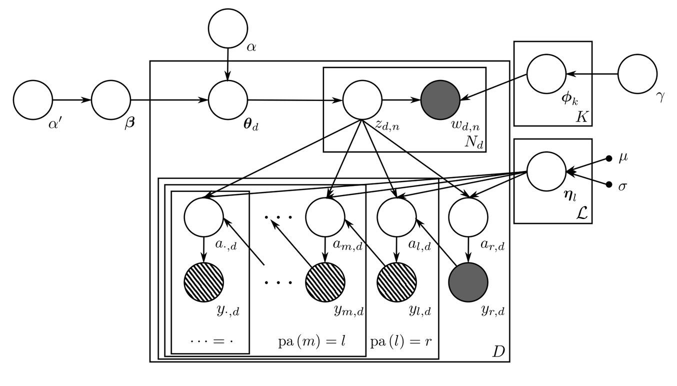


Figure 1: HSLDA graphical model

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     – Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     – Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     – Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{\mathbf{z}}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     – Set the response variable
       $$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing.

In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [8]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [8]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [8] we find

$$\begin{aligned} p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \\ \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \end{aligned} \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [12]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [11].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha\boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [10, 9, 13, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [**?**]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[8] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[9] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[10] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[11] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[12] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[13] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

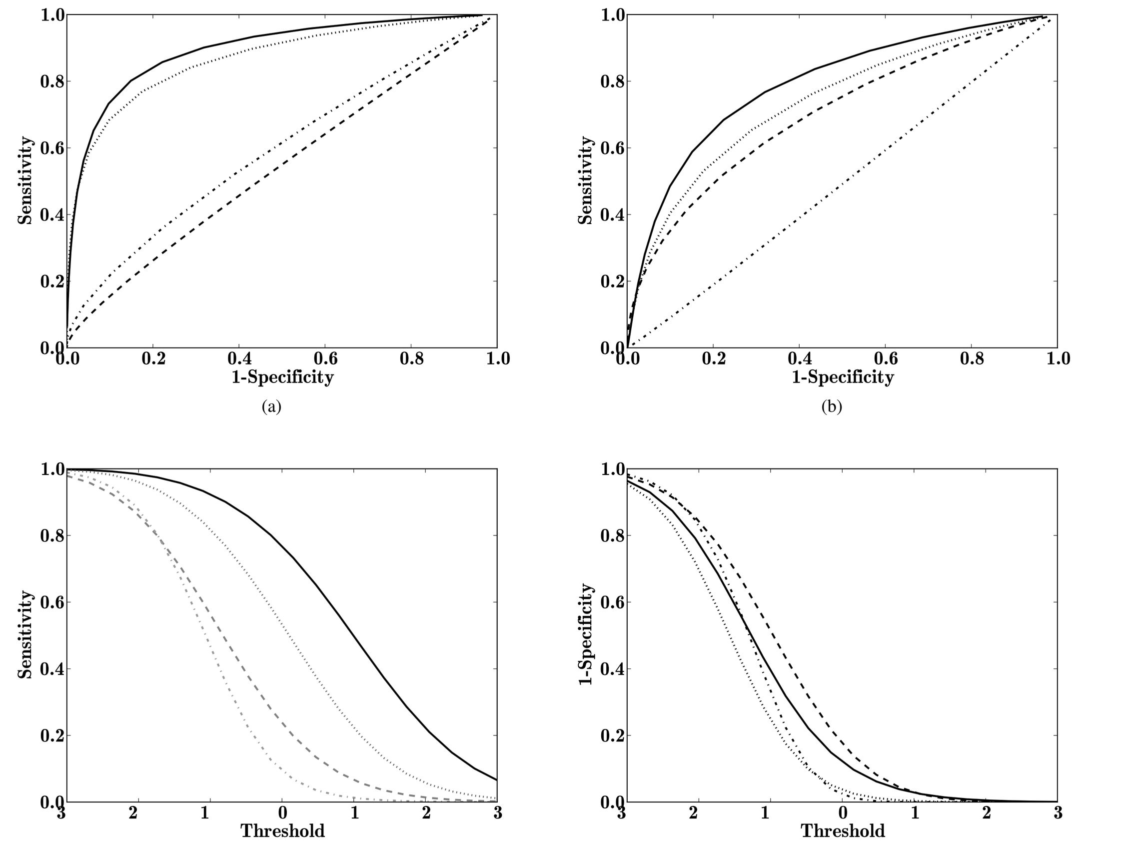


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

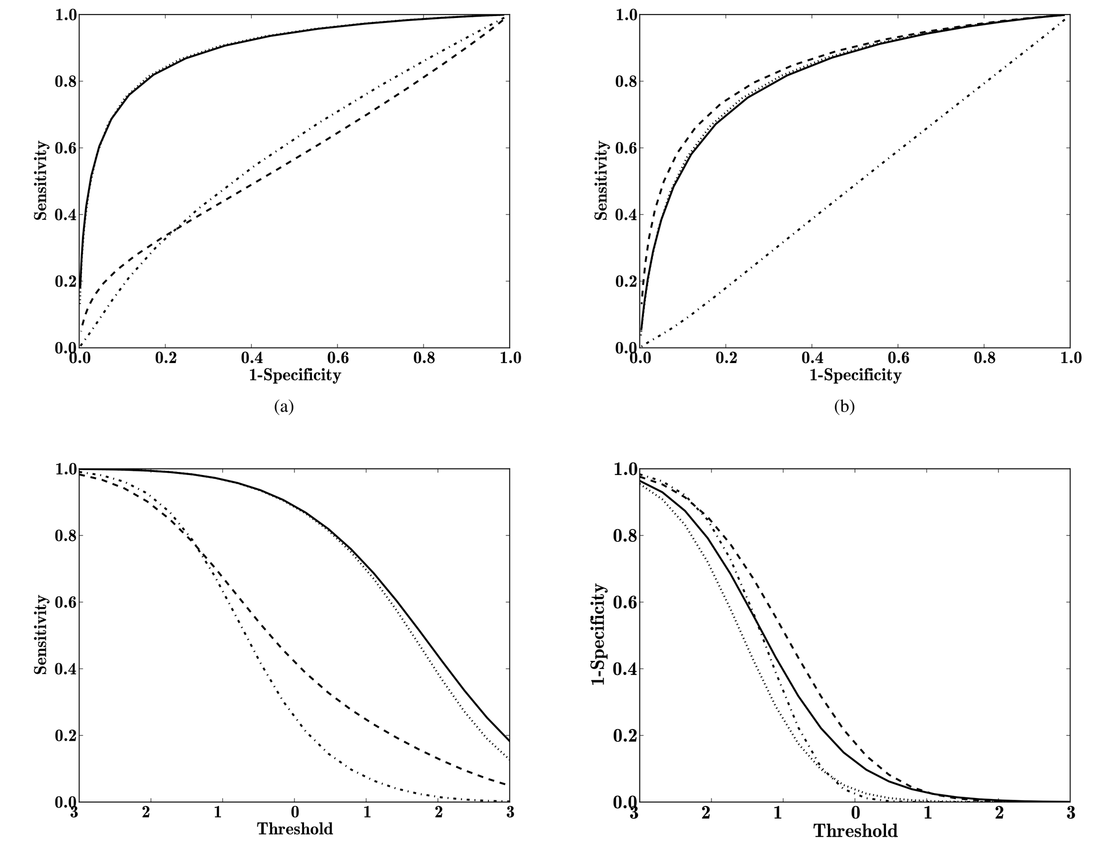


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [4], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

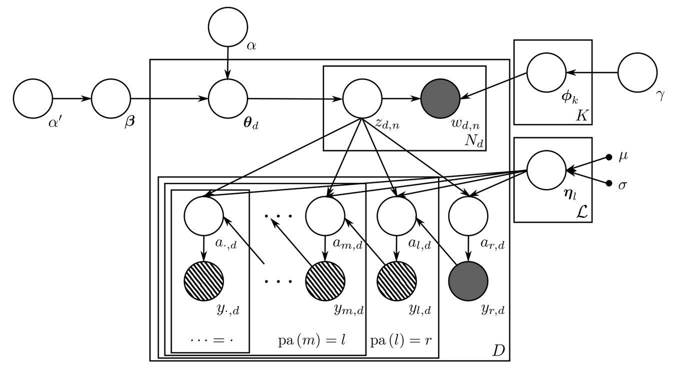


Figure 1: HSLDA graphical model

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \quad (1)$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \quad (2)$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \quad (3)$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \quad (4)$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (5)$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (6)$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{d,k})} s\left(n_{d,k}, m\right) \left(\alpha\beta_k\right)^m \quad (7)$$

where $s(n,m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [**?**], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [**?**].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [**?**]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [**? ? ?**]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [**? ? ? ? ?**].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

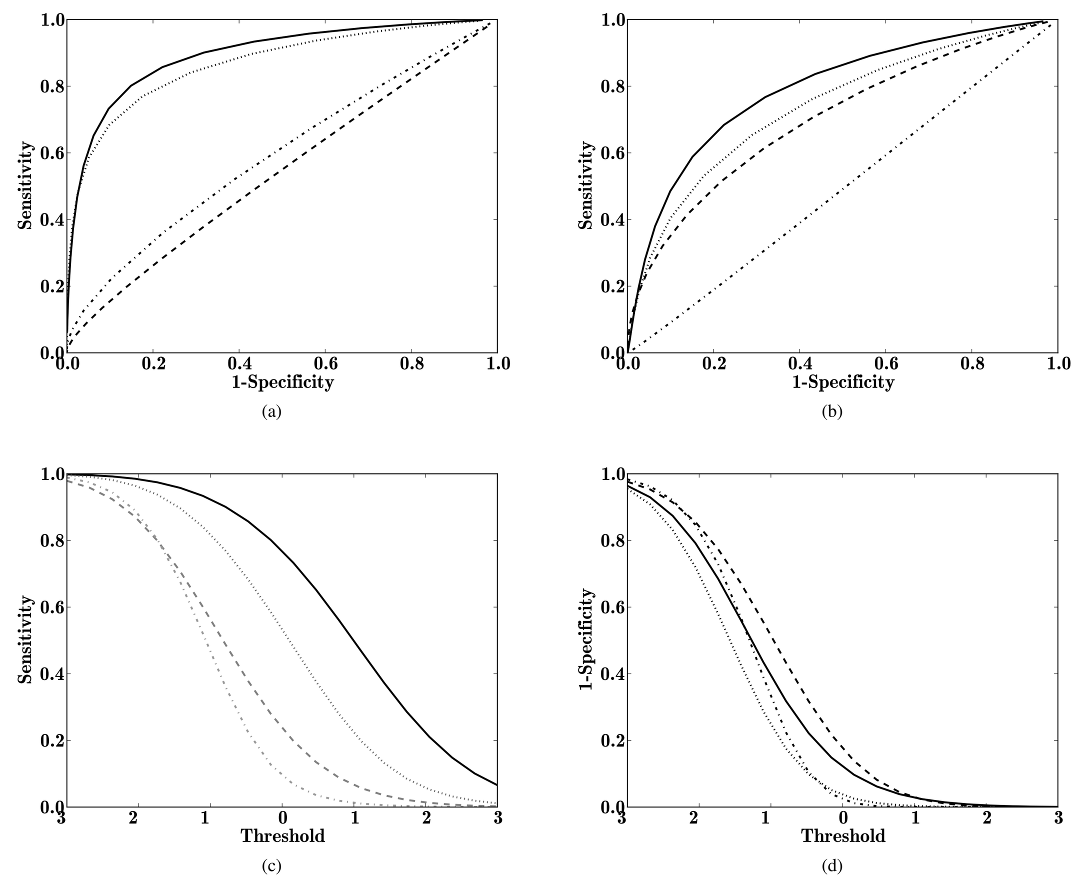


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

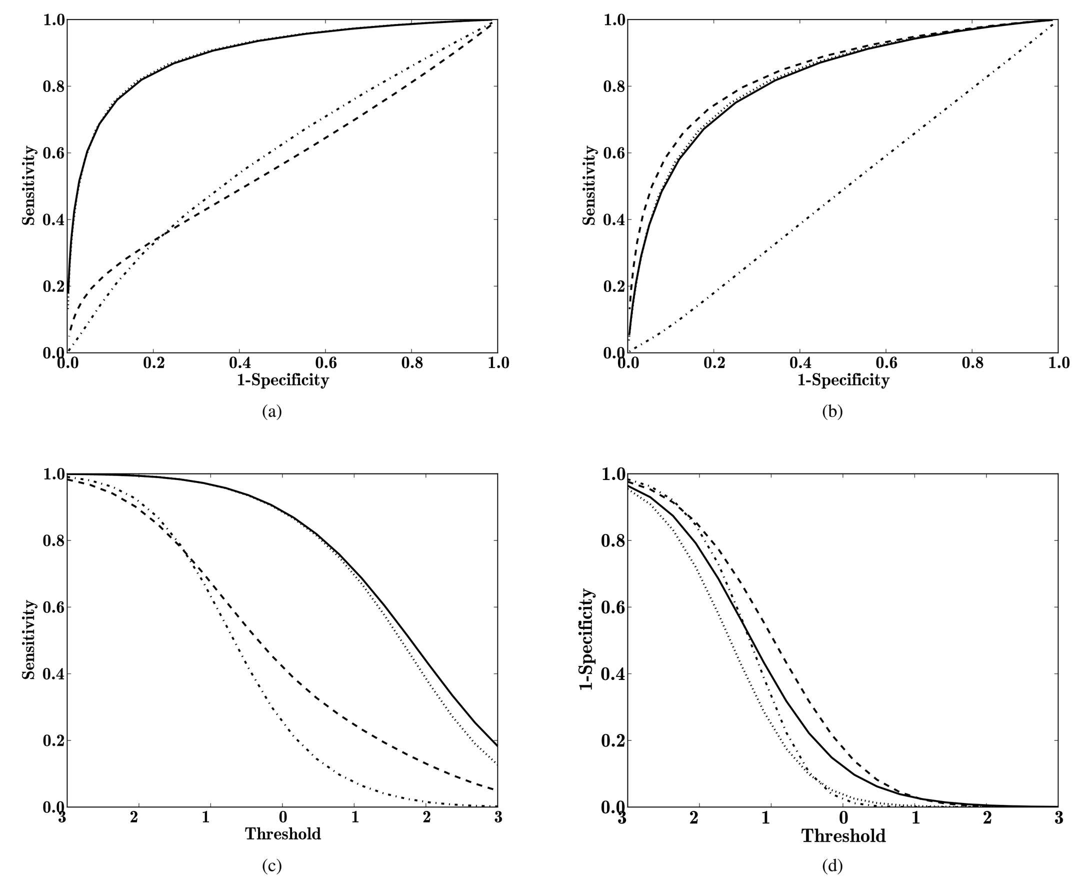


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noemie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [1] as available from [4]), and patient treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better performance than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

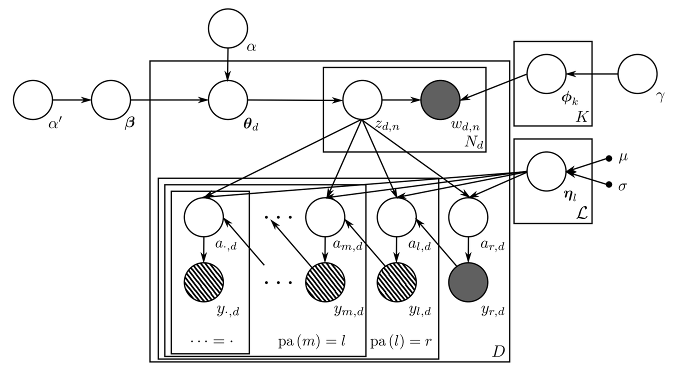


Figure 1: HSLDA graphical model

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$

       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

       $$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing.

In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \setminus \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}$, α, α' and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \setminus z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_l \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

In the section we describe the application of HSLDA for prediction in two hierarchically structured domains. Firstly, we describe using discharge summaries to predict diagnoses, encoded as ICD-9 codes. Discharge summaries are documents that are authored by clinicians to summarize the course of a hospitalization. ICD-9 codes are used mainly for billing purposes to indicate the conditions for which a patient was treated. Secondly, we describe using Amazon.com product descriptions to predict product categories.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Our data set was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patient's chief complaint, diagnostic findings, therapy administered, patient's response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code representing "Pneumonia due to adenovirus" is a child of the code representing "Viral pneumonia" where the former is a type of the latter. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary.

The text of the discharge summaries were pre-processed such that each document would be represented as counts over a 10,000 word vocabulary. The Natural Language Toolkit was used to tokenize the text. A vocabulary was identified by first sorting terms based on a global term frequency-inverse document frequency measure. The top 10,000 words which were not identifying in some way (a name, place, or identifying number) were selected for inclusion in the vocabulary.

For each hospitalization there are usually several ICD-9 codes assigned for billing purposes. These codes are known to be quite specific but not very sensitive [8]. Regardless of that fact, this is one of the only sources for information on patient diagnoses aside from the free text.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. This study was approved by the Institutional Review Board.

5.1.2 Product Category Prediction

Data for these experiments were obtained partially from the Stanford Network Analysis Platform (SNAP) Amazon product metadata dataset [4] and partially directly from the the Amazon.com website [1]. The product ID's and categorizations were obtained from the SNAP dataset and the product descriptions were obtained directly from the website.

We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine". Each product was labeled with multiple categories.

The vocabulary for this experiment was created by including the most frequent 30K words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

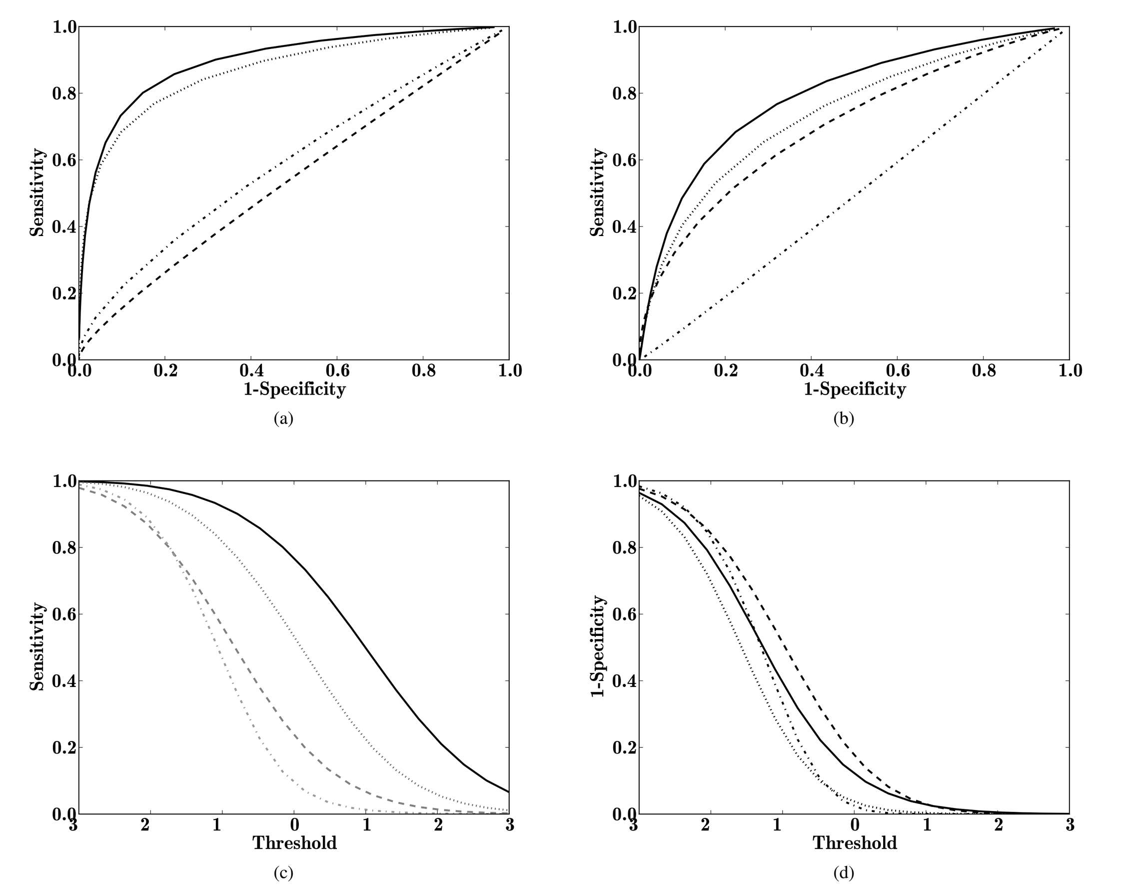


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

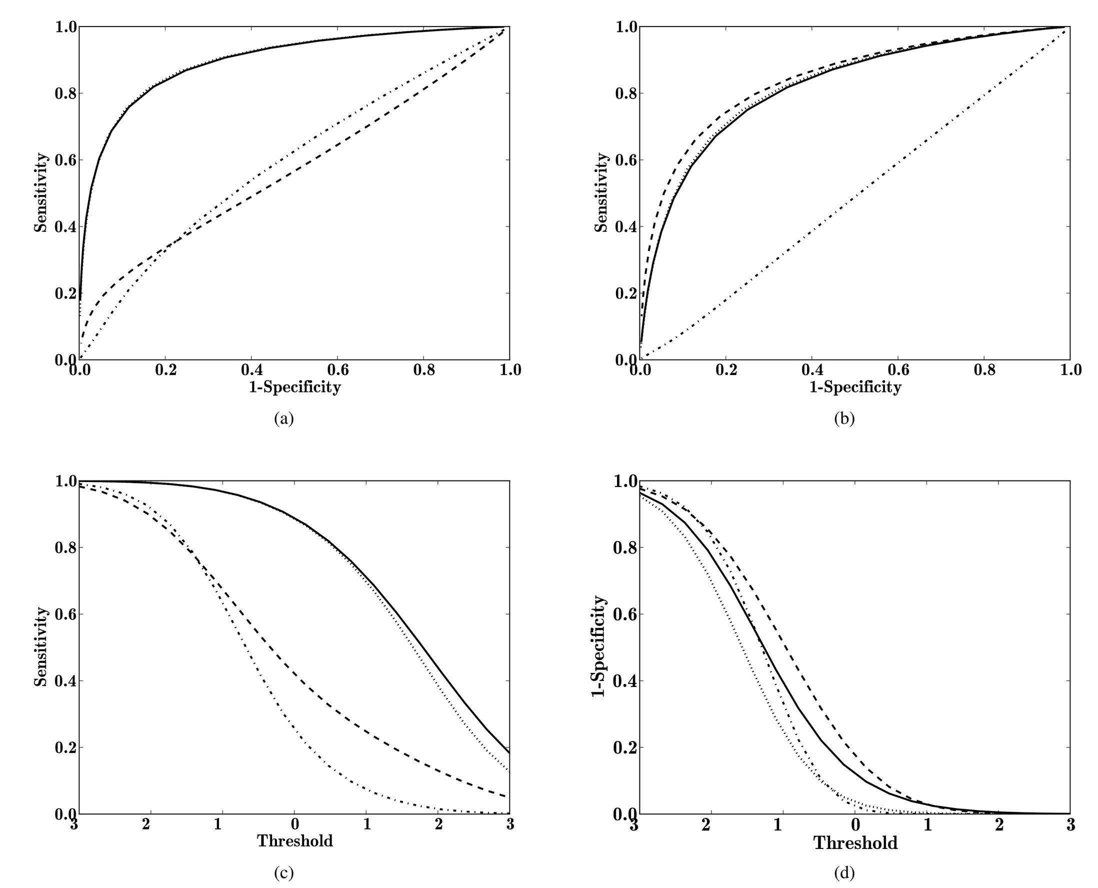


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noémie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also have a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

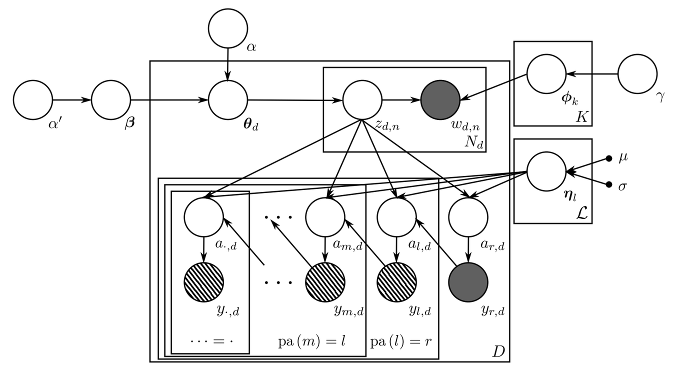


Figure 1: HSLDA graphical model

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, μ, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is shown in Figure 1.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T\boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1}\sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L}\backslash\mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [12]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l\in\mathcal{L}}$, $\mathbf{a} = \{a_{l',d}\}_{l'\in\mathcal{L}_d, d=1,\ldots,D}$, $\boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [12]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [12] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T\boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}} \tag{3}$$
$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [?]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T\bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d}y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [21]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [20].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

3.1.5 $p\left(\alpha\right), p\left(\alpha'\right), p\left(\gamma\right)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

There have been many models that incorporate both latent models of text and some form of supervision [17, 13, 22, 6]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[] Amazon, Inc. http://www.amazon.com/, 2011.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

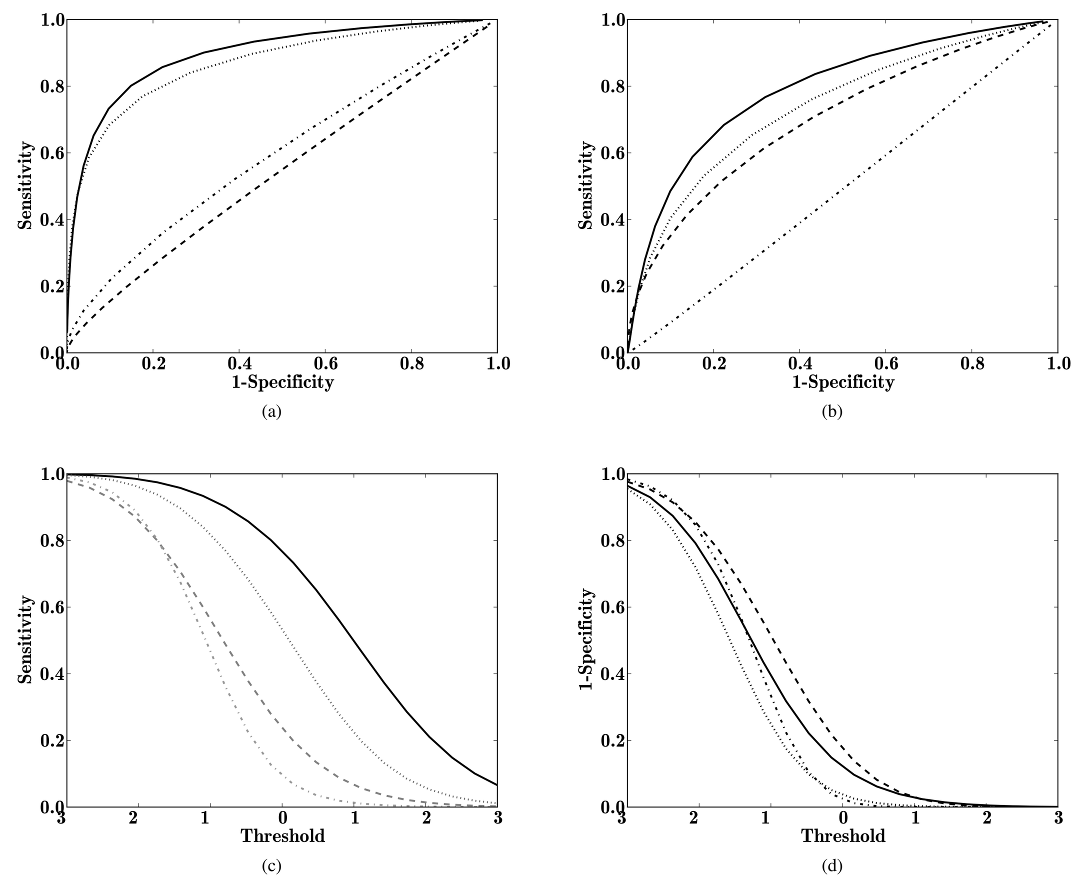


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

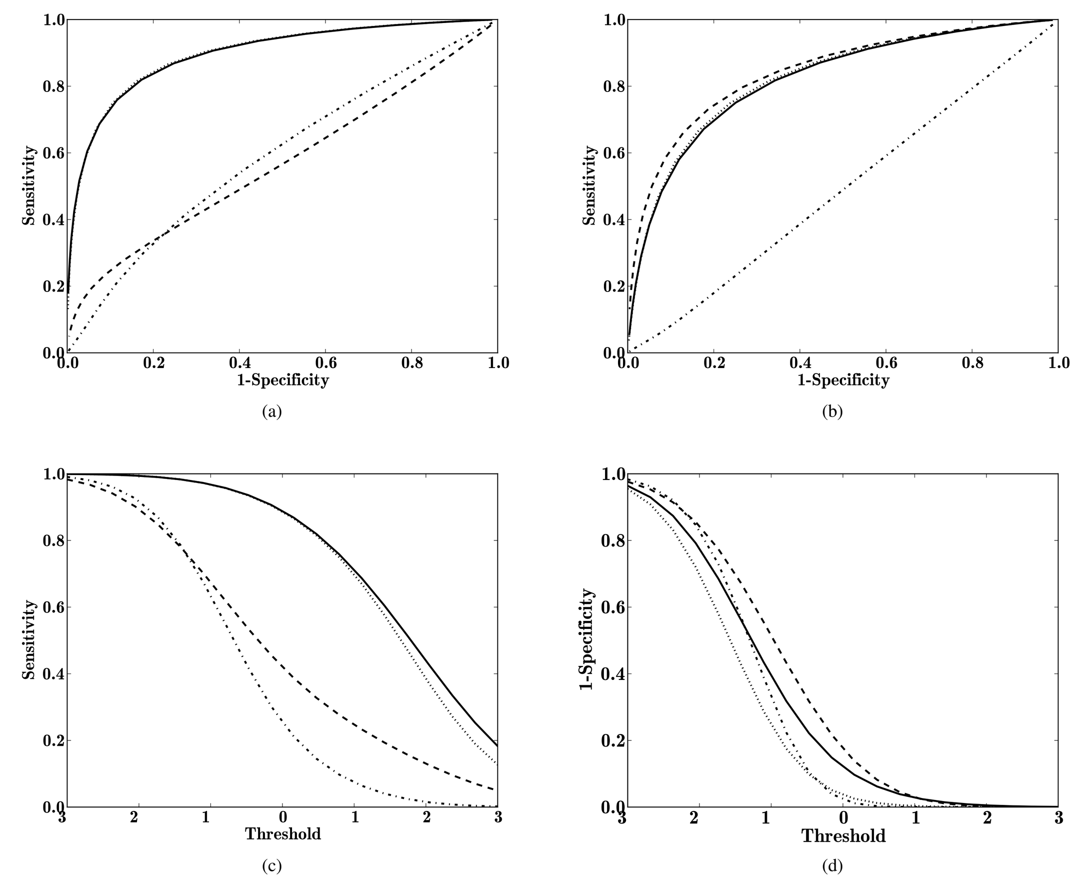


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [4], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [5].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as

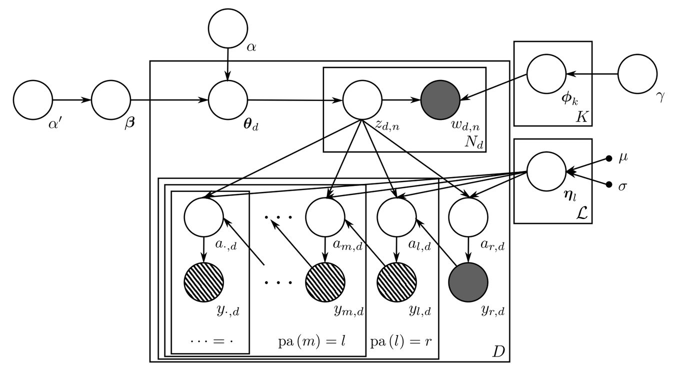


Figure 1: HSLDA graphical model

not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure **??** where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure **??**. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (**??**) and the regression coefficients (**??**)(**??**) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [9]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [9]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [9] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c_{v,d}^{k,-(n,d')}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 2, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$
$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta})$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [13]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [12].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been many models that incorporate both latent models of text and some form of supervision [11, 10, 14, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [**?**], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [**?**].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [**?**]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [**? ? ?**]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [**? ? ? ? ?**].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[2] http://www.nltk.org

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[9] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[11] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[13] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[14] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

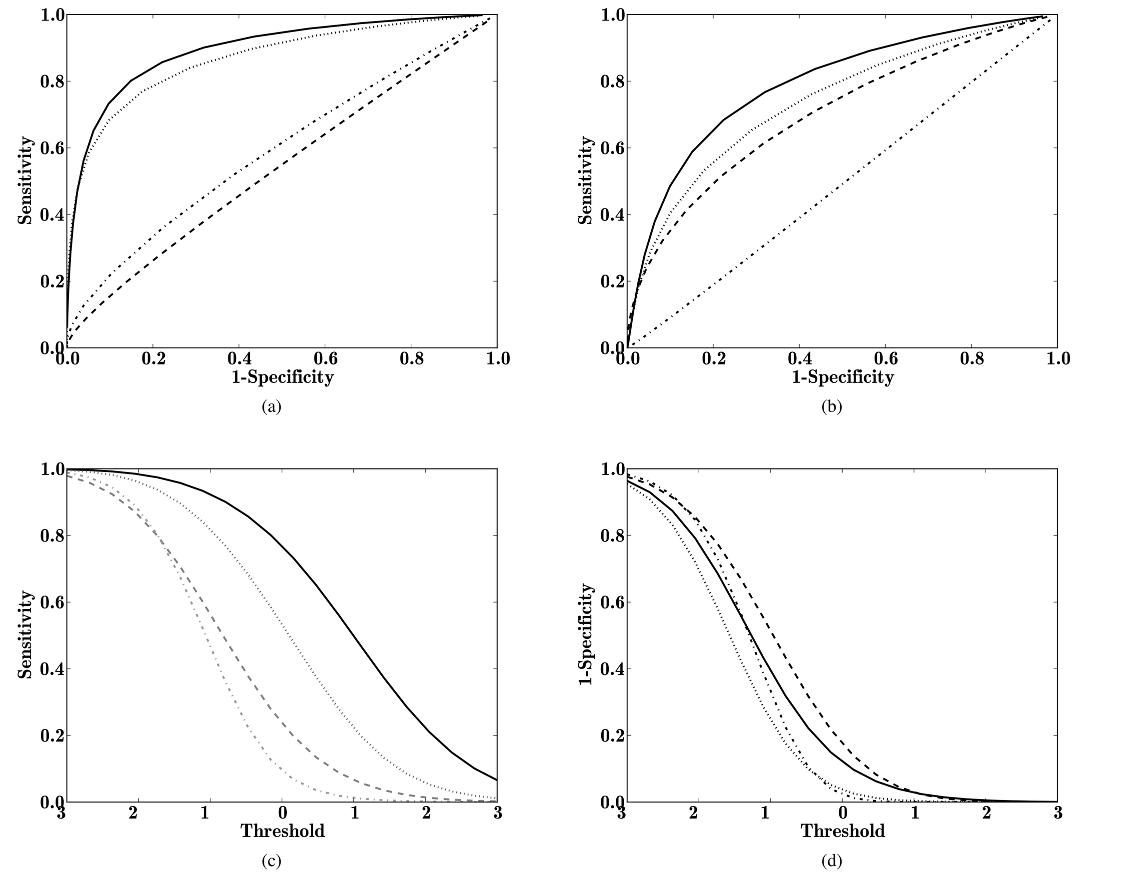


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

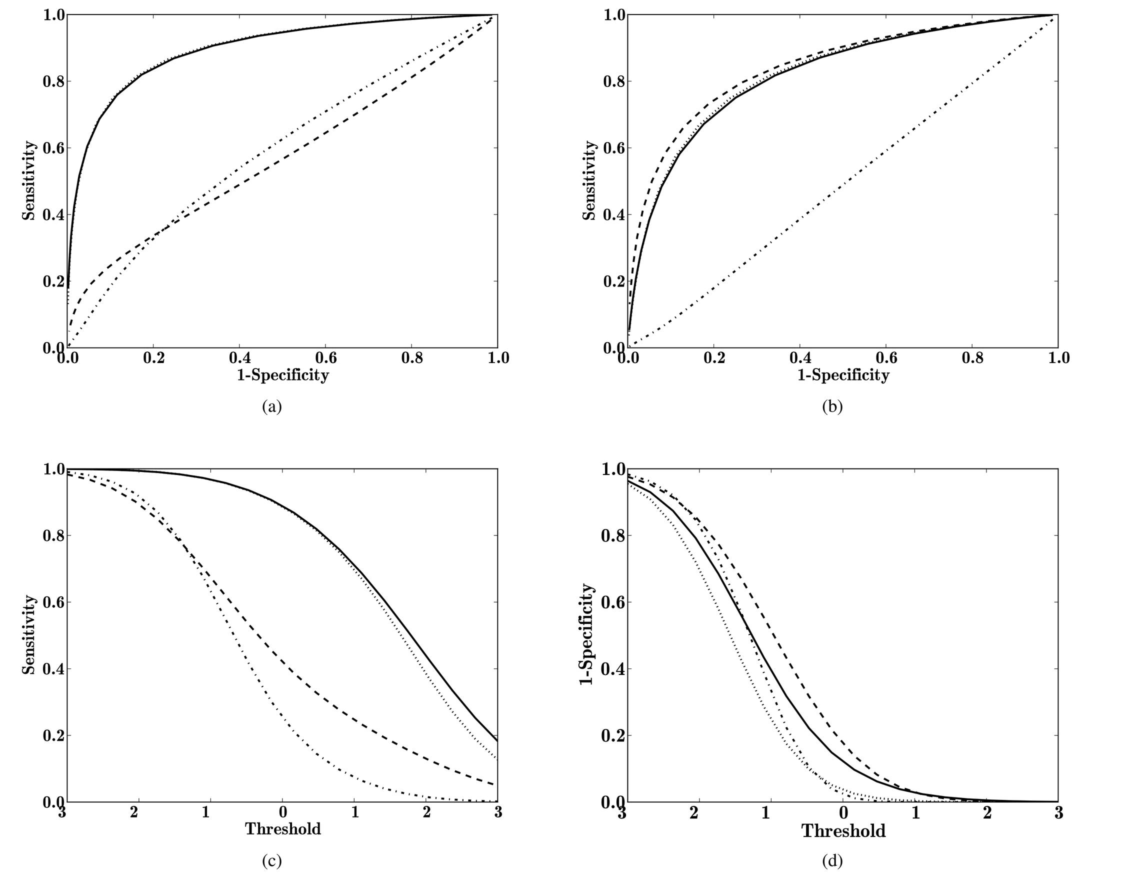


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte   Nicholas Bartlett   Noémie Elhadad   Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generally modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labels in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

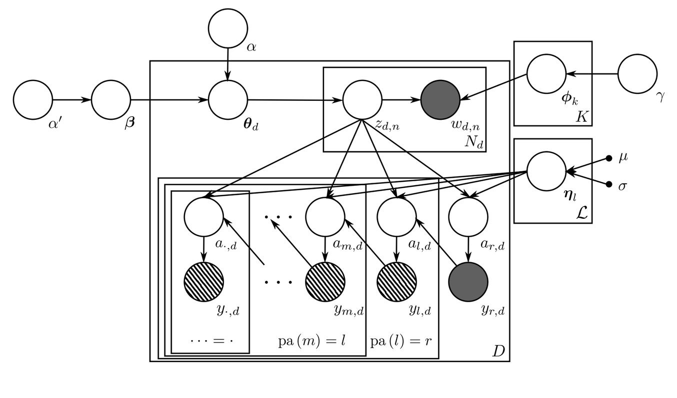


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (**??**) and the regression coefficients (**??**) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [12], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d')}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

is also a standard probit regression result [**?**].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to ADLER [21]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [20]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are ADLER

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind. ADLER check this. I suspect that some of the n's need some summing.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

There have been many models that incorporate both latent models of text and some form of supervision [17, 13, 22, 6]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where the lack of a link did not truly indicate absence. In hierarchically labeled data, negative labels are uncommon and the lack of a label in the hierarchy is not equivalent to a negative label. Therefore, as in the work of Chang and Blei, we employ regularization to account for the lack of negative labels. This will be discussed further in 2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

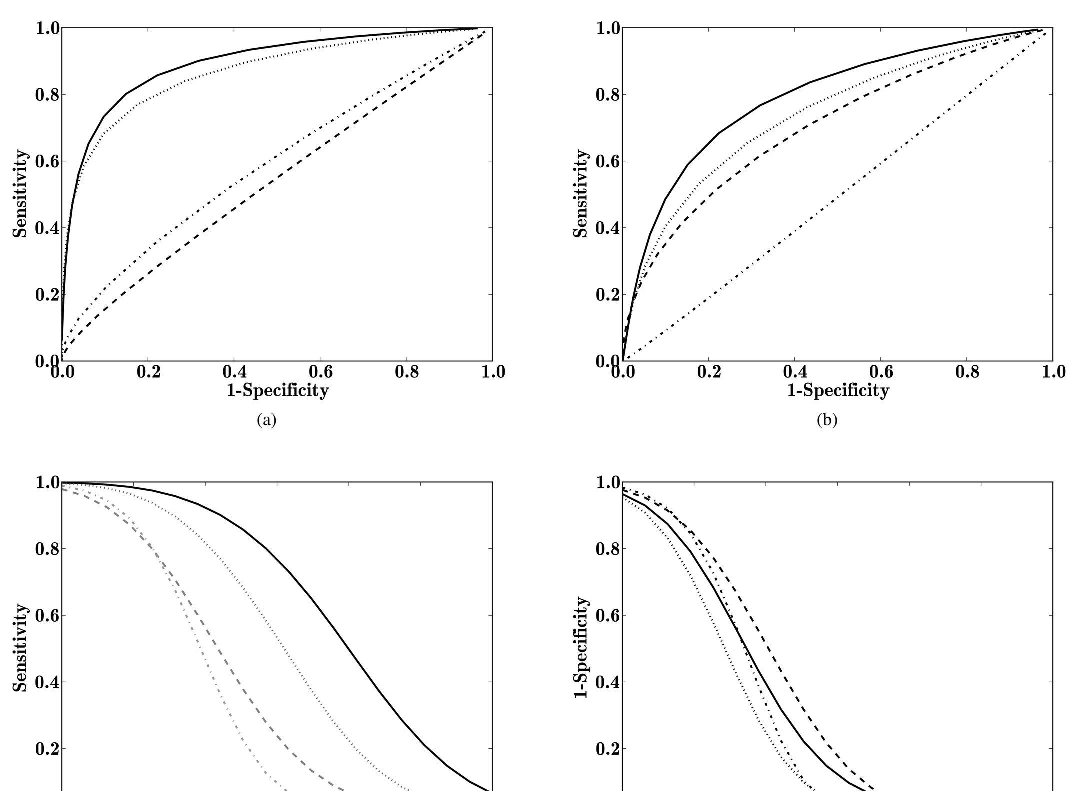


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

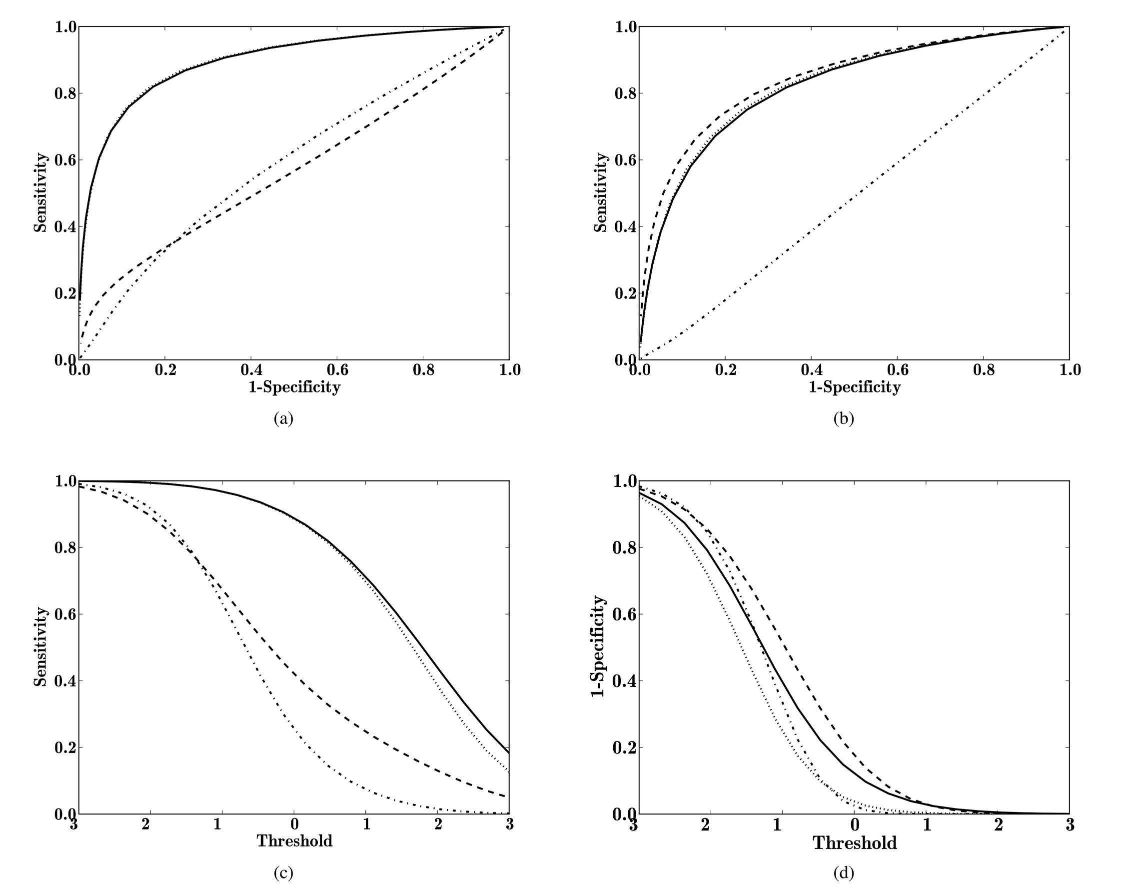


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[14] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[15] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[16] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[19] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

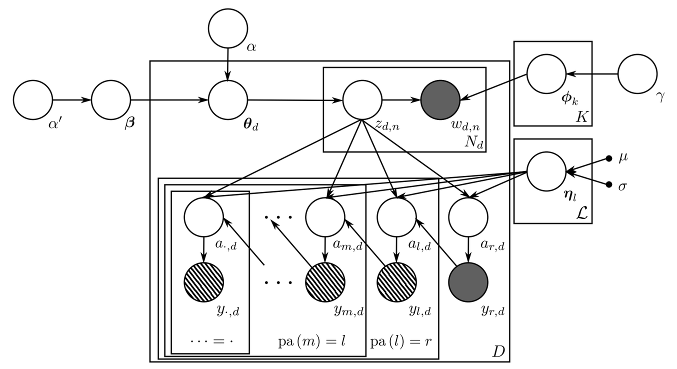


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**?** **?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [12], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c^{k,-(n,d')}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

is also a standard probit regression result [**?**].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to ADLER [21]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [20]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are ADLER

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma(\alpha \boldsymbol{\beta}_k + n_{d,k})} s(n_{d,k}, m) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind. ADLER check this. I suspect that some of the n's need some summing.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating supervision in the form of an observed exponential family response variable per document. As a result, sLDA infers topics such that the model predicts the response variable while improing word likelihood. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

Other models that incorporate both and some form of supervision include LabeledLDA[17], DiscLDA[13], models applied to computer vision and document networks[22, 6].

2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

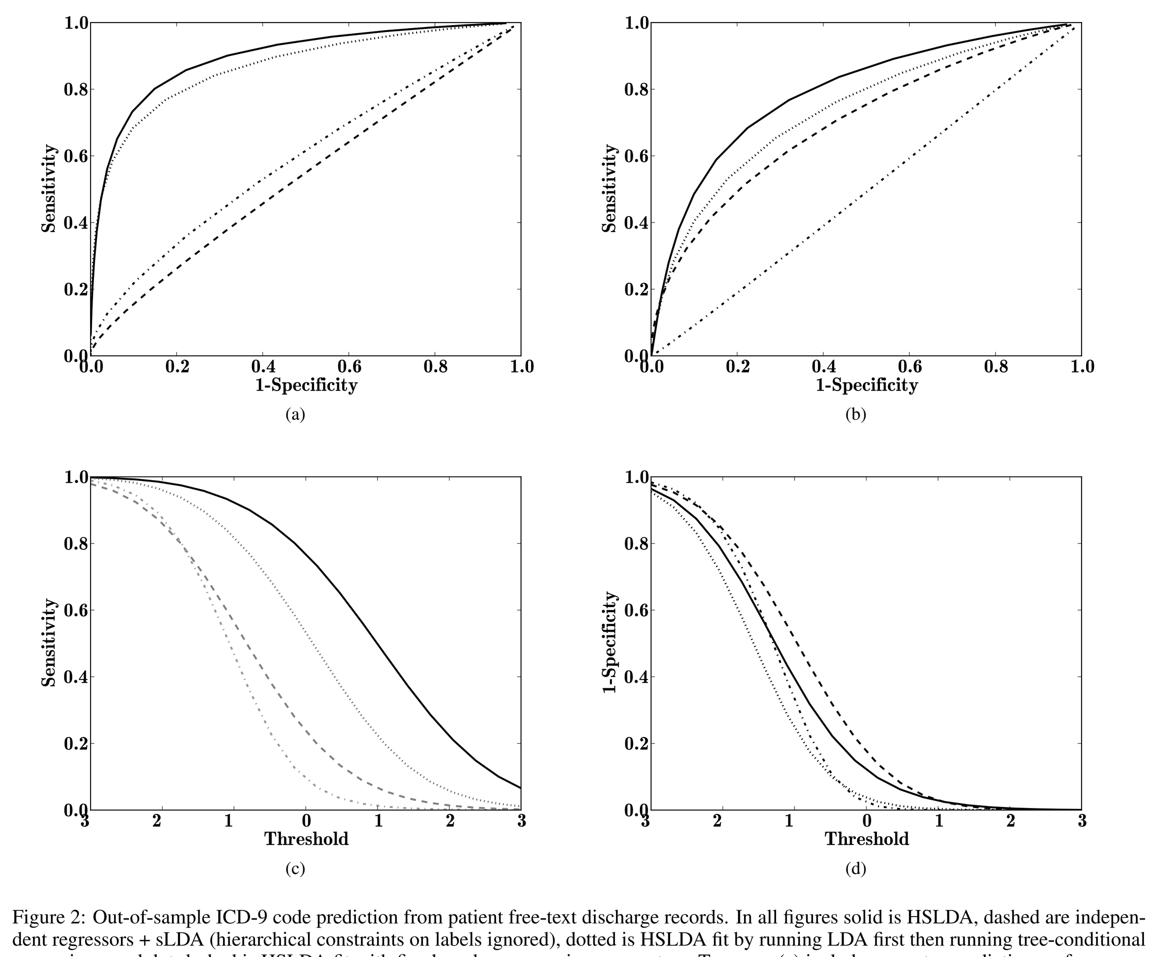


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

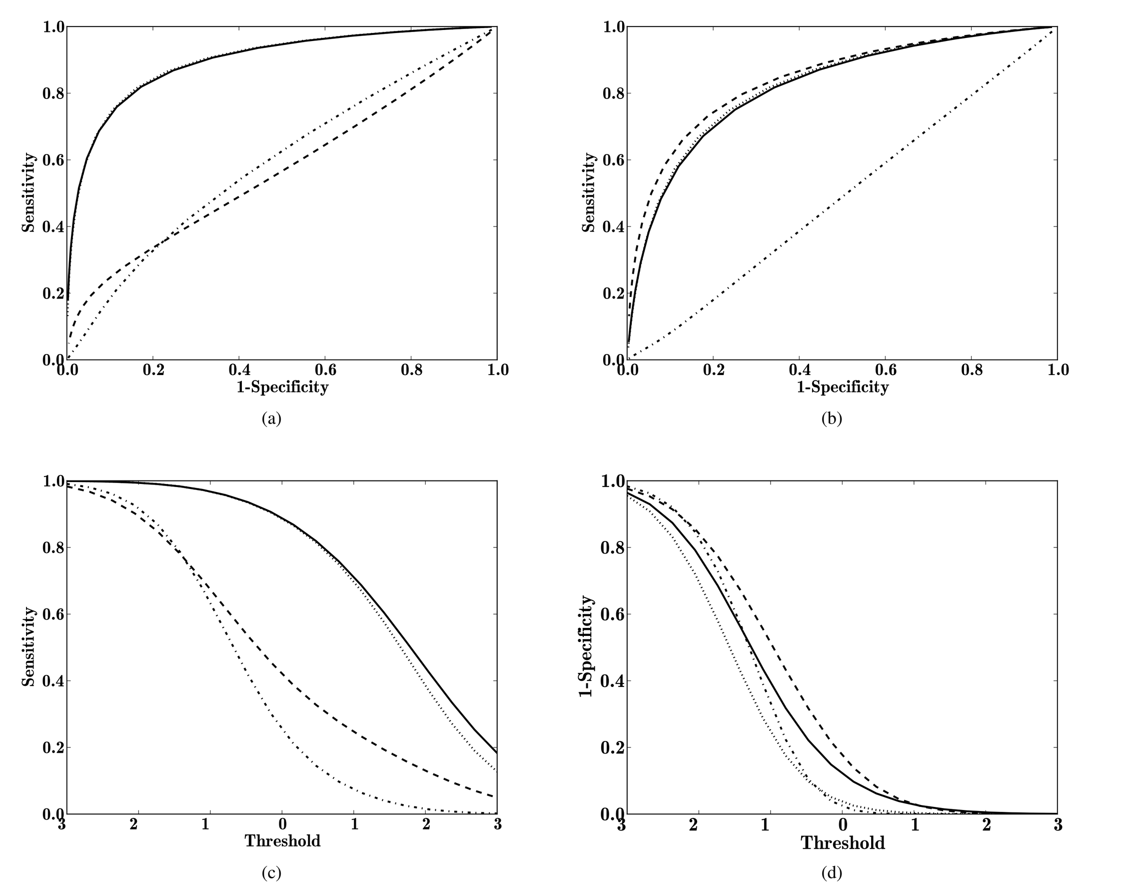


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[14] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[15] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[16] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[19] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

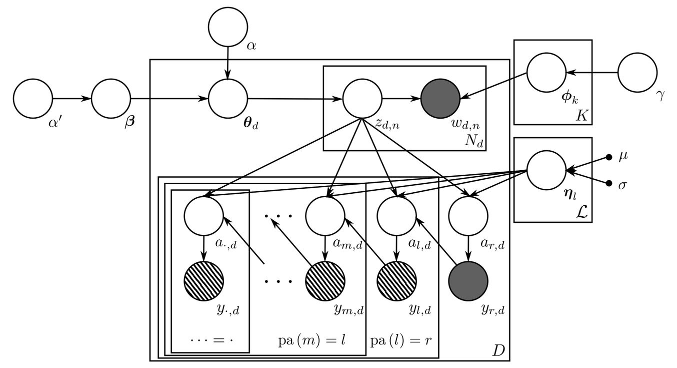


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**?** **?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [12], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d')}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

is also a standard probit regression result [**?**].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [21]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [20]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{\cdot,d}^k\right)} s\left(c_{\cdot,d}^k, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{5}$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating supervision in the form of an observed exponential family response variable per document. As a result, sLDA infers topics such that the model predicts the response variable while improing word likelihood. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

Other models that incorporate both and some form of supervision include LabeledLDA[17], DiscLDA[13], models applied to computer vision and document networks[22, 6].

2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

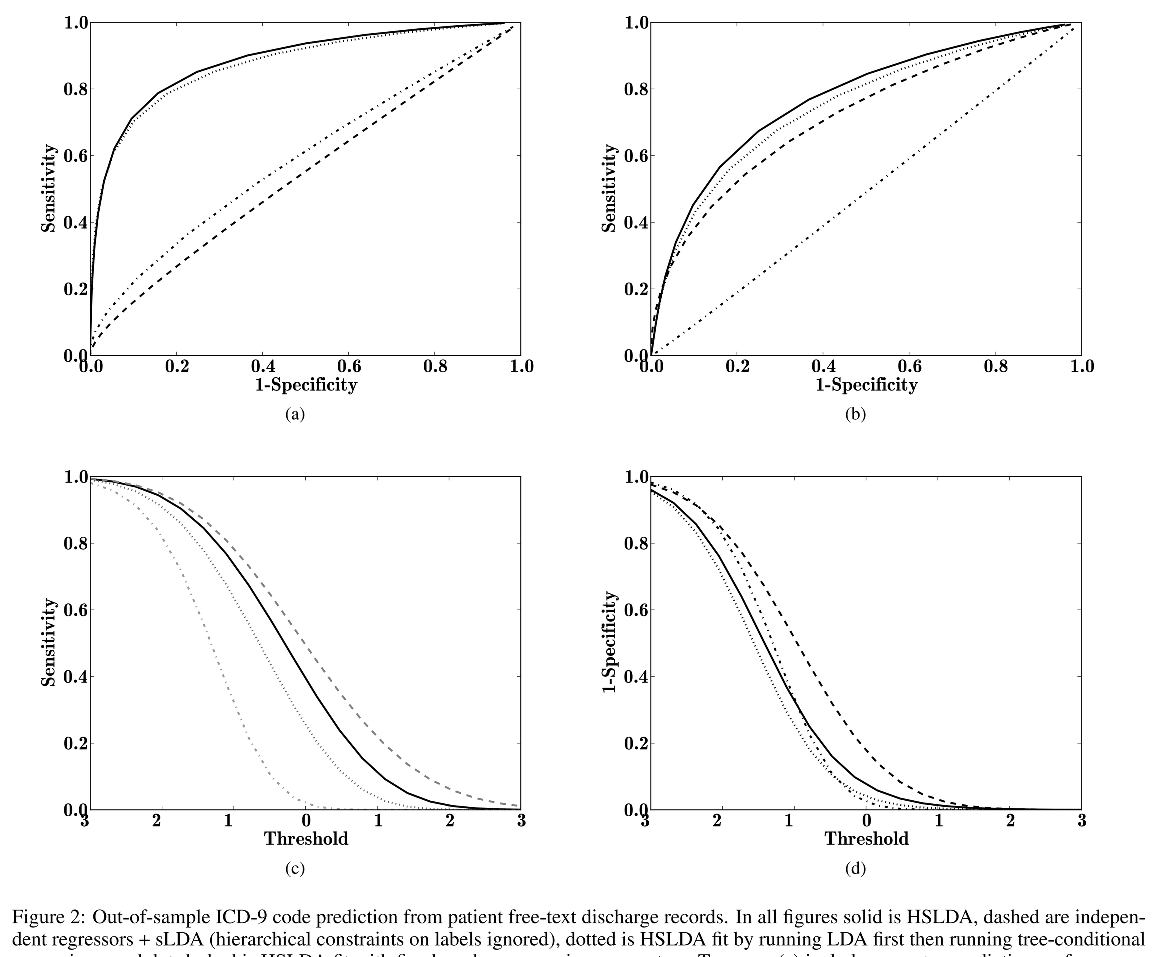


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

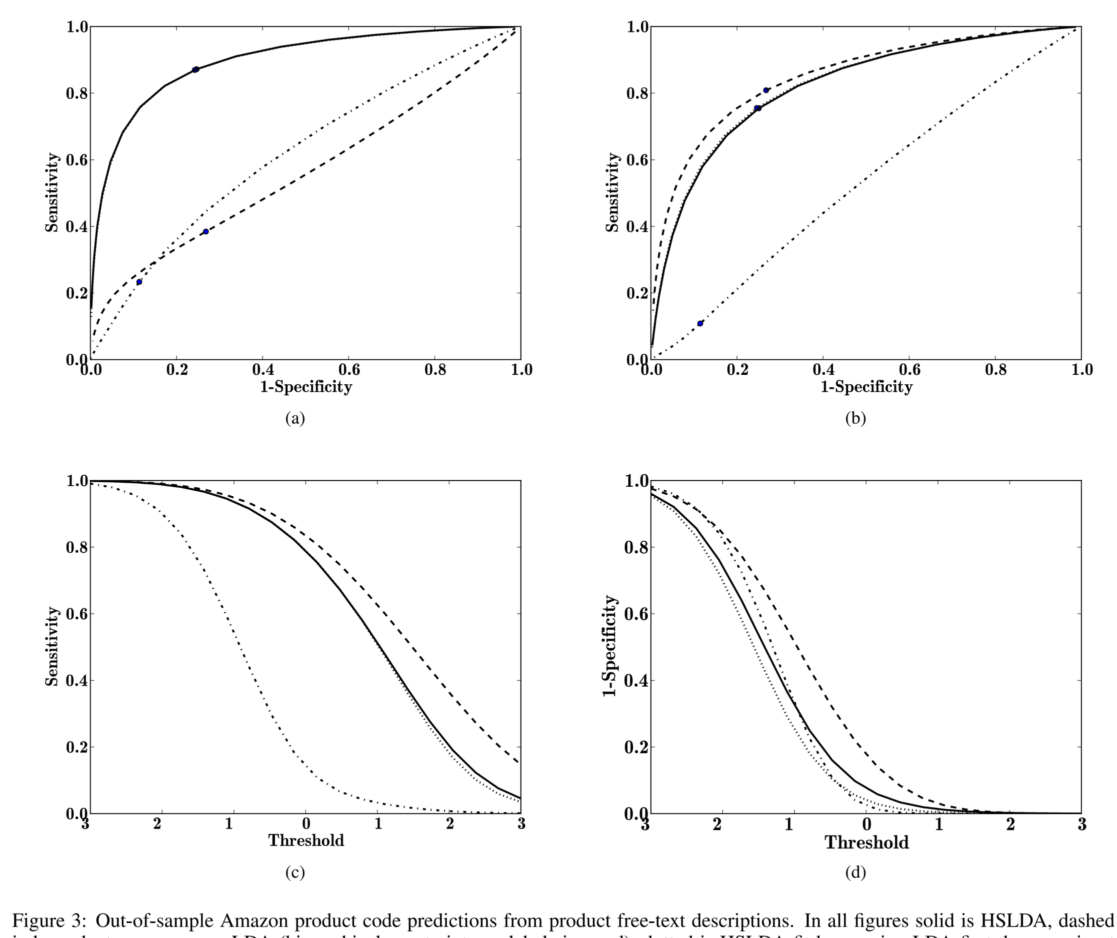


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

6 Results

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[14] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[15] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[16] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[18] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[19] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

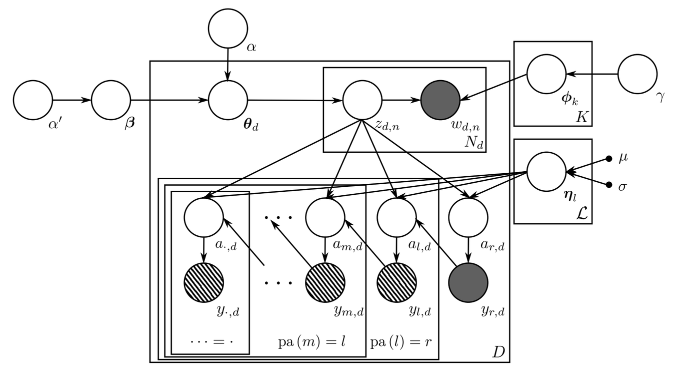


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [12], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto$$
$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d')}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [?]; the specific form of the update is a standard result from Bayesian normal linear regression [?]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

is also a standard probit regression result [?].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [21]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [20]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{\cdot,d}^k\right)} s\left(c_{\cdot,d}^k, m\right)\left(\alpha \boldsymbol{\beta}_k\right)^m \tag{5}$$

$s\left(n, m\right)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating supervision in the form of an observed exponential family response variable per document. As a result, sLDA infers topics such that the model predicts the response variable while improing word likelihood. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

Other models that incorporate both and some form of supervision include LabeledLDA[17], DiscLDA[13], models applied to computer vision and document networks[22, 6].

2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

6 Results

HSLDA	sLDA with Independent Regressors
children, animation, animated, kids, songs, story, family, new, fun,video, friends, music, live, classic, adventure, minute, world, magic,animals, song, musical, voiced, feature, along, baby, ages, magical, action, favorite, special, tale, little, child, holiday, help, voice, toys, original, learn, characters, computer, sing, like, adults, parents, version, old, first, including, young,	children, kids, animated, animation, songs, video, friends, fun, animals, family, story, new, little, music, like, minute, along, musical, song, adventure, help, classic, baby, child, ages, favorite, sing, learn, holiday, magical, tale, world, special, parents, live, way, day, boy, make, old, magic, adventures, home, show, gang, animal, adults, toys many, action
comedy, funny, comic, hilarious, series, gags, jokes, slapstick, satire, laugh, stars, films, comedies, first, comedian, funniest, show, involving, characters, cast, perfect, mob, performance, play, writer, television, satirical, lines, hilariously, boss, routine, silent, funnier, comedic, chemistry, comedians, talk, routines, brothers, girlfriend, shorts, mouse, whose, cinema, screwball, hit, humor, brilliant, gag, new,	comedy, show, play, characters, sketches, funny, mob, performance, people, screwball, cast, best, including, stars, sketch, girlfriend, hilariously, charming, ensemble, comic, character, star, writer, hit, stage, host, classic, romantic, mutants, boss, new, played, together, choice, like, karaoke, daughter, vehicle, news, splitting, actress, get, story, hilarious, wacky, featuring, could, side, comedies, comedi
thriller, horror, murder, crime, killer, police, mystery, detective, cop, one, dead, suspense, plot, case, blood, drug, dark, mysterious, death, revenge, violence, town, gang, turns, prison, noir, violent, deadly, stars, young, woman, nbsp, body, vampire, night, murders, kill, criminal, action, becomes, wife, cops, gangster, sex, victims, victim, brutal, terror, evil, murdered	horror, film, vampire, blood, one, dead, supernatural, killer, evil, thriller, gore, monster, young, night, terror, mysterious, director, movie, films, cult, creepy, terrifying, original, dark, chilling, genre, atmosphere, victims, classic, murder, haunted, town, mystery, story, body, ghost, tale, family, murders, castle, house, nightmare, budget, eerie, death, doctor, victim, later, bloody, effects
workout, body, yoga, video, moves, dance, minute, program, fitness, easy, poses, muscles, routine, exercises, learn, strength, movements, practice, help, step, w, workouts, minutes, techniques, flexibility, get, exercise, one, instructor, first, series, time, instruction, use, follow, training, breathing, two, basic, back, ll, muscle, fun, warm, steps, technique, balance, beginners, designed, using	workout, body, yoga, video, moves, dance, minute, program, fitness easy, poses, muscles, exercises, learn, routine, strength, movements, minutes, get, help, practice, step, series, techniques, workouts, time, flexibility, exercise, instructor, instruction, one, two, fun, training, use, first, back, follow, steps, breathing, muscle, basic, using, ll, beginners, three, warm, balance, technique, designed

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

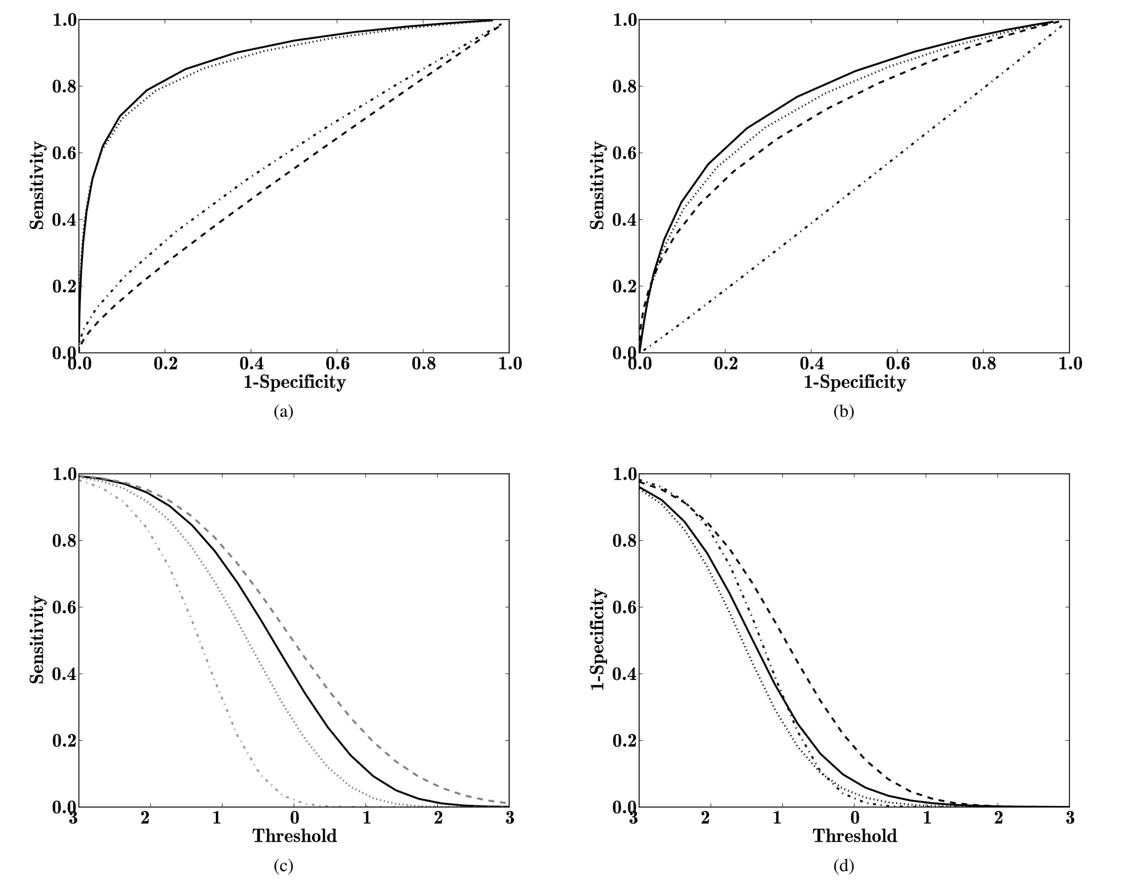


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

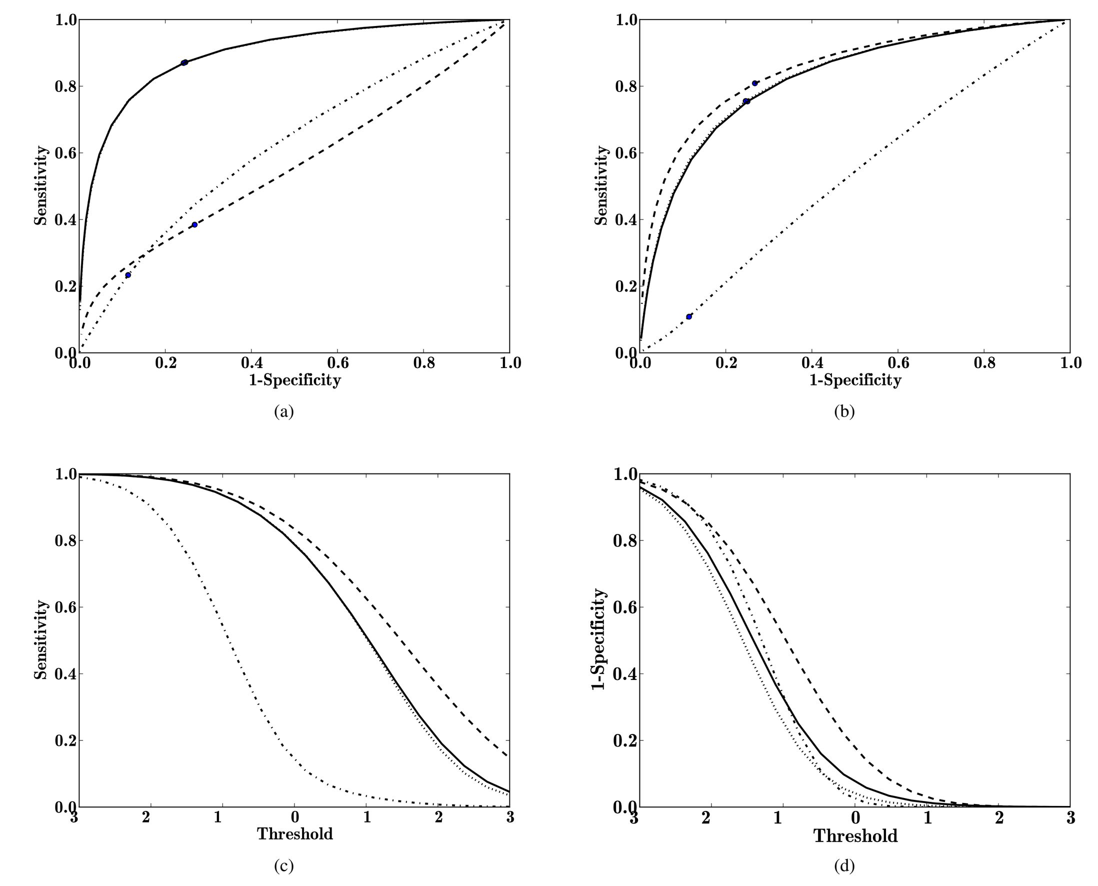


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[14] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[15] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[16] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[18] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[19] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte Nicholas Bartlett Noémie Elhadad Frank Wood
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurrence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a trees with a single root $r \in \mathcal{L}$. Each document has a response $y_{l,d} \in \{-1, 1\}$ to every label which indicates whether the label applies to document d or not.

The is-a hierarchical constraint is a hard constraint that document d has a positive response to label l_2 then it will also have a positive response to label l_1. Conversely, if document d has a negative response to label l_1 then it will also has a negative response to l_2. To capture this hierarchical structure we model the labeling of documents using a generative cascade of conditional probit regression models.

Each document is assigned a response of either -1 or 1 for at least one, but potentially many labels in $\mathcal{L}$. The label, l, for a document, d, will be used interchangeably to refer to the observed response of document d to label l.

The fixed parameters of the model are the number of topics K, the number of unique words in the vocabulary V, the number of documents D, as well as the mean, $\boldsymbol{\mu}$, and the standard deviation, σ, used in a normal prior distribution. The hyper-parameters α', α, and γ are weight

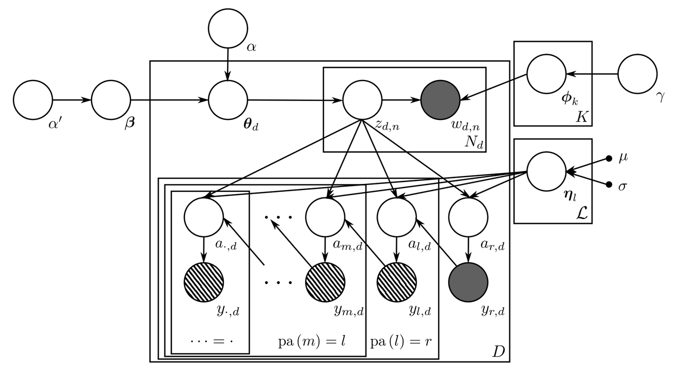


Figure 1: HSLDA graphical model

parameters for Dirichlet prior distributions. We will denote the K-dimensional Dirichlet distribution as $\mathrm{Dir}_K(\cdot)$ and the K dimensional normal distribution as $\mathcal{N}_K(\cdot)$.

We will now describe the stochastic generative process which defines our model. The graphical model is show in Figure **??**.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a regression coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$, where $\mathbf{I}_K$ is the K dimensional identity matrix
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - For each label $l \in \mathcal{L}$:
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l)} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l)} = -1 \end{cases}$
       where $\bar{z}_d = N_d^{-1} \sum_{n=1}^{N} z_{n,d}$
     - Set the response variable

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \text{ and } y_{\mathrm{pa}(l),d} = 1 \\ -1 & \text{otherwise} \end{cases}$$

This type of generative model is known as a probit regression model. Probit regression models are a type of discriminative probabilistic model similar to logistic regression. However, instead of using the logistic sigmoid as the link function, the probit regression model uses the CDF for a standard normal distribution - the inverse of which is known as the probit. In this case, the regression is conditional on the parents according to the constraints of the labeling hierarchy. The latent variables $a_{l,d}$ utilized here are also known as an auxiliary variables because the are introduced to make exact Gibbs sampling possible and are not of primary interest.

Given that negative labels are uncommon and that the absence of a label is not equivalent to a negative label, we apply an informative prior to the regression parameters, $\boldsymbol{\beta}_{\mathcal{L}}$, in the form of a negative prior that encodes a bias towards being truly negative in the absence of a label.

3 Inference

In the Bayesian approach to statistical modeling, the primary task of inference is to find the posterior distribution over the unobserved parameters of the model. However, it is often possible and desirable to integrate over certain variables in the model, also known as collapsing. In our model, it will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. It is straightforward to integrate out the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ from the full generative model. We can also integrate out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ as in Griffiths and Steyvers [13]. Therefore, in our model the latent variables are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

The posterior distribution we seek cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior.

3.1 Gibbs Sampler

We derive a collapsed Gibbs sampler for this model by considering the individual conditional probability distributions for each of the unobserved variables. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

3.1.1 $p(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma)$

First we consider the conditional distribution of the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$. The conditional distribution does not include $\theta_{1:D}$ and $\phi_{1:K}$ because they have been integrated out as in the collapsed Gibbs sampler [13]. The conditional distribution of $z_{n,d}$ is proportional to the joint distribution of its markov blanket.

$$p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \tag{1}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in [13] we find

$$p\left(z_{n,d} = k \mid \mathbf{z}_{-(n,d)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\}. \tag{2}$$

Here, $c^{k,-(n,d')}_{v,d}$ represents the number of words of type v in document d assigned to topic k omitting the n^{th} word of document d'. The notation $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Given Equation 1, $p\left(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right)$ can be sampled through enumeration.

3.1.2 $p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma)$

We now consider the conditional distribution of the regression coefficients $\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$ such that

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}} \tag{3}$$

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}_i \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right). \tag{4}$$

This is a standard result from normal Bayesian linear regression [**?**]. Here, $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$.

3.1.3 $p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right)$

The auxiliary variables $a_{l,d}$ must be sampled for documents $d = 1, \ldots, D$ and $l \in \mathcal{L}_d$. The conditional posterior distribution of $a_{l,d}$ is the truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{5}$$

This conditional distribution can be sampled using an inverse CDF method.

3.1.4 $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha\right)$

In our model, we place a hierarchical Dirichlet prior over topic assignments. This flexible distribution allows for an asymmetric prior over document level distributions over topics [22]. This prior shares many features with the hierarchical Dirichlet process and inference over this distribution proceeds in a very similar fashion.

Posterior inference is performed using the "direct assignment" method of Teh et al. [21].

$$\boldsymbol{\beta} \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{6}$$

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + n_{d,k}\right)} s\left(n_{d,k}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{7}$$

where $s(n, m)$ represents stirling numbers of the first kind.

3.1.5 $p(\alpha), p(\alpha'), p(\gamma)$

The hyperparameters α, α', and γ are given broad $\mathrm{Gamma}(1, 1000)$ prior distributions and sampled via the Metropolis-Hastings algorithm.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document

labeling, often taking the form of a single numerical or categorical label. Examples of labels include rating associated with an online eview, grades for an essay, and number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

There have been several models that incorporate both latent models of text and some form of supervision [18, 14, 23, 7]. One set of models that are particularly relevant to HSLDA are Chang and Blei's hierarchical models for document networks (Relational Topic Models). In that family of models, they encountered a similar scenario where an unselected label does not always indicate absence. In hierarchical labels, this phenomenon is even more pervasive – there are no explicit negative labels, but it is also unclear how to treat the parents of selected labels. Like in the work of Chang and Blei, we employ regularization to account for the lack of negative labeling. In our experiments, we look at the impact of assigning positive and negative instances to the ancestors of selected labels.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [9], and sometimes make mistakes [11].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [8, 12, 10]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [15, 19, 17, 20, 16].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

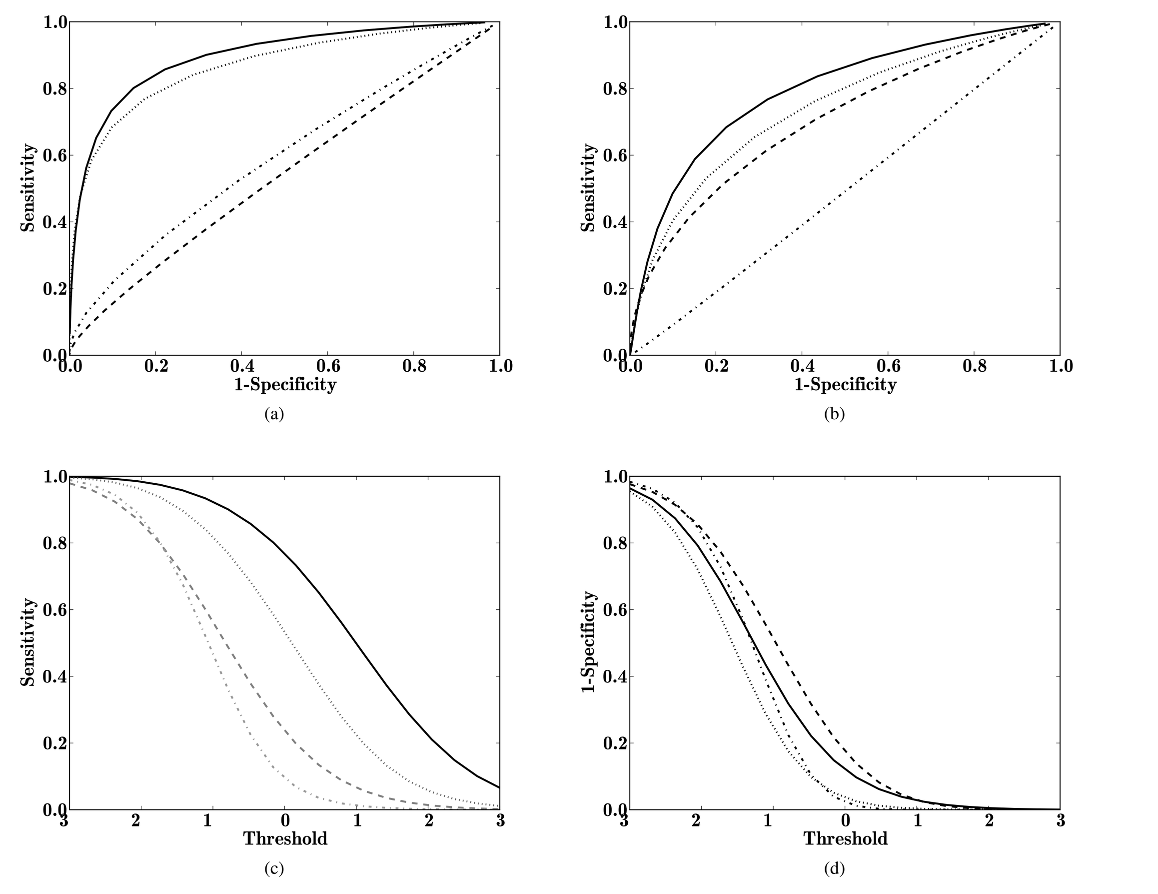


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by performing LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

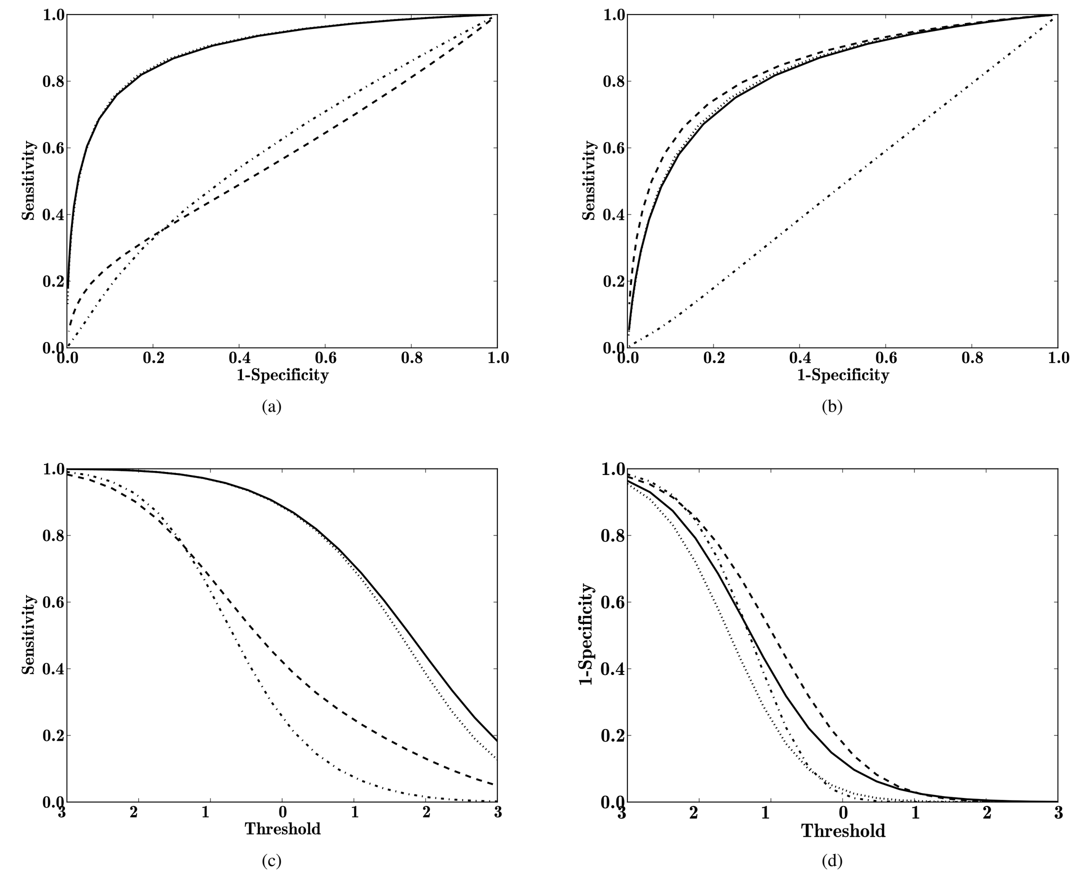


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[10] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[11] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[12] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[13] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[14] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[15] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[17] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
`{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu`

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure **??** where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

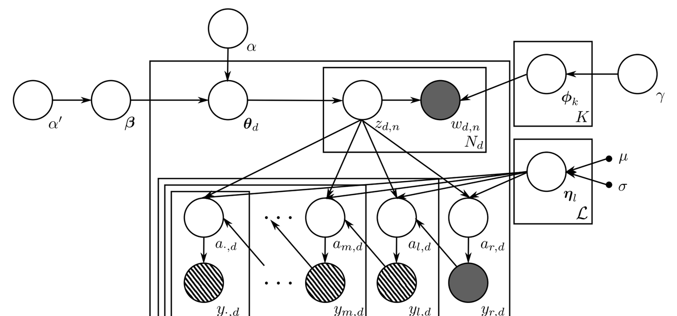


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure **??**. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (**??**) and the regression coefficients (**??**) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [13], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c^{k,-(n,d')}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regressor [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

is also a standard probit regression result [4].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [22]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [21]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha\beta_k)^m \tag{5}$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include rating associated with an online eview, grades for an essay, and number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate both and some form of supervision include LabeledLDA[18], DiscLDA[14], models applied to computer vision and document networks[23, 7].

2.

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Diagnosis Prediction

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [9], and sometimes make mistakes [11].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [8, 12, 10]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [15, 19, 17, 20, 16].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Category Prediction

In this experiment, we look at product descriptions and their categorizations according to a product hierarchy. Product ID's and categorizations were obtained from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were crawled from the amazon.com website directly. We were able to deduce the structure of the hierarchy for the Amazon.com products directly since all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. We will refer to the three comparison models as the sLDA model, the separate HSLDA model, and the random HSLDA model, respectively. These models were chosen as to highlight performance in absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

For each dataset, a held out set of 1,000 documents and accompanying labels were used for evaluation. The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

6 Results

HSLDA	sLDA with Independent Regressors
children, animation, animated, kids, songs, story, family, new, fun,video, friends, music, live, classic, adventure, minute, world, magic,animals, song, musical, voiced, feature, along, baby, ages, magical, action, favorite, special, tale, little, child, holiday, help, voice, toys, original, learn, characters, computer, sing, like, adults, parents, version, old, first, including, young,	children, kids, animated, animation, songs, video, friends, fun, animals, family, story, new, little, music, like, minute, along, musical, song, adventure, help, classic, baby, child, ages, favorite, sing, learn, holiday, magical, tale, world, special, parents, live, way, day, boy, make, old, magic, adventures, home, show, gang, animal, adults, toys, many, action
comedy, funny, comic, hilarious, series, gags, jokes, slapstick, satire, laugh, stars, films, comedies, first, comedian, funniest, show, involving, characters, cast, perfect, mob, performance, play, writer, television, satirical, lines, hilariously, boss, routine, silent, funnier, comedic, chemistry, comedians, talk, routines, brothers, girlfriend, shorts, mouse, whose, cinema, screwball, hit, humor, brilliant, gag, new,	comedy, show, play, characters, sketches, funny, mob, performance, people, screwball, cast, best, including, stars, sketch, girlfriend, hilariously, charming, ensemble, comic, character, star, writer, hit, stage, host, classic, romantic, mutants, boss, new, played, together, choice, like, karaoke, daughter, vehicle, news, splitting, actress, get, story, hilarious, wacky, featuring, could, side, comedies, comedi
thriller, horror, murder, crime, killer, police, mystery, detective, cop, one, dead, suspense, plot, case, blood, drug, dark, mysterious, death, revenge, violence, town, gang, turns, prison, noir, violent, deadly, stars, young, woman, nbsp, body, vampire, night, murders, kill, criminal, action, becomes, wife, cops, gangster, sex, victims, victim, brutal, terror, evil, murdered	horror, film, vampire, blood, one, dead, supernatural, killer, evil, thriller, gore, monster, young, night, terror, mysterious, director, movie, films, cult, creepy, terrifying, original, dark, chilling, genre, atmosphere, victims, classic, murder, haunted, town, mystery, story, body, ghost, tale, family, murders, castle, house, nightmare, budget, eerie, death, doctor, victim, later, bloody, effects
workout, body, yoga, video, moves, dance, minute, program, fitness, easy, poses, muscles, routine, exercises, learn, strength, movements, practice, help, step, w, workouts, minutes, techniques, flexibility, get, exercise, one, instructor, first, series, time, instruction, use, follow, training, breathing, two, basic, back, ll, muscle, fun, warm, steps, technique, balance, beginners, designed, using	workout, body, yoga, video, moves, dance, minute, program, fitness easy, poses, muscles, exercises, learn, routine, strength, movements, minutes, get, help, practice, step, series, techniques, workouts, time, flexibility, exercise, instructor, instruction, one, two, fun, training, use, first, back, follow, steps, breathing, muscle, basic, using, ll, beginners, three, warm, balance, technique, designed

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[9] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[10] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[11] Mehdrad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[12] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[13] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[14] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

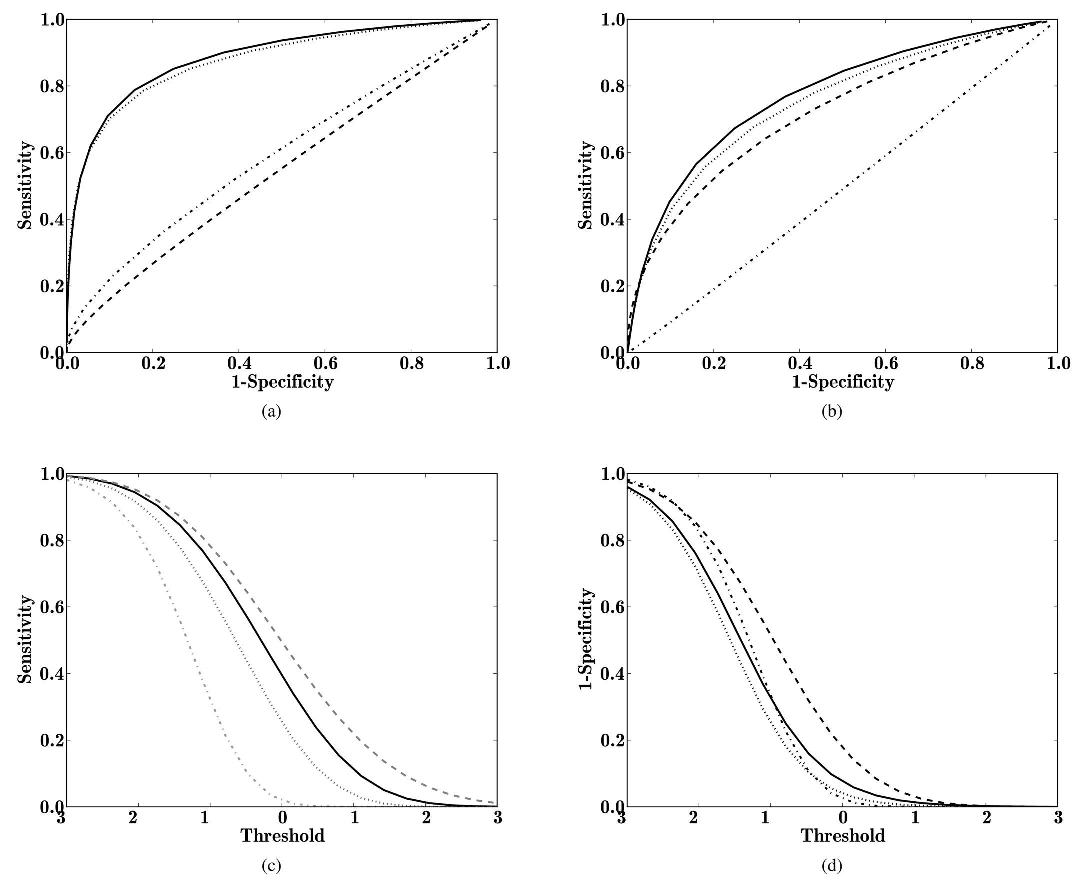


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

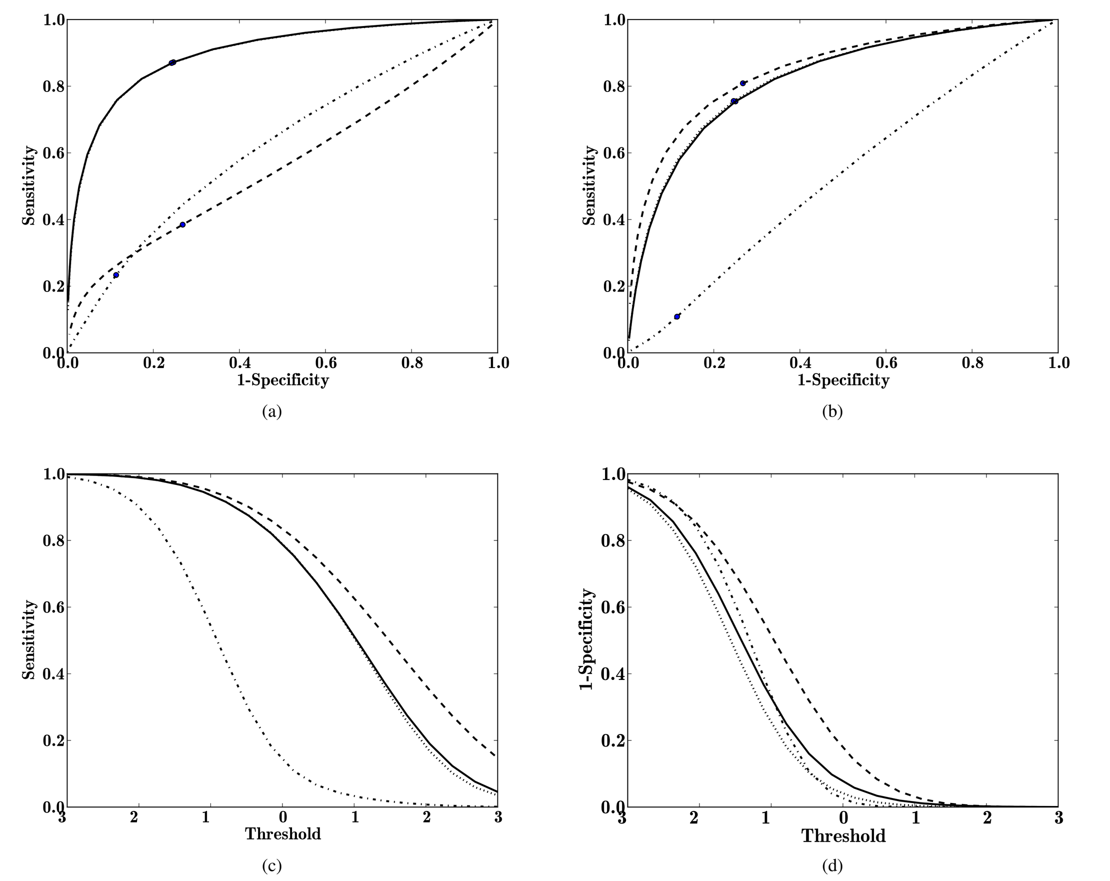


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[15] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[17] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply-labeled bag-of-word data. Examples of such data include web pages and their placement in link directories, product descriptions and placement(s) in product hierarchies, and free-text clinical records and diagnosis codes assigned to them. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data is also of interest. We find that using the signal from hierarchical labels substantially improves out-of-sample label prediction in comparison to other models that don't utilize the structure of the labels. We demonstrate HSLDA on large-scale data from medical document labeling and retail product categorization tasks. We show improved label prediction performance and evidence that the learned topics also improve.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs [3], and patient records and their associated billing codes. In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

In this work we extend supervised latent Dirichlet allocation (sLDA) [4] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [5] augmented with per document "supervision"; often taking the form of a single numerical or categorical "label." More generally this supervision is just extra per document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as having been conditionally drawn from some distribution that depends on the document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [4].

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is structured as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), Section 3 details a sampling approach to inference in HSLDA, Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

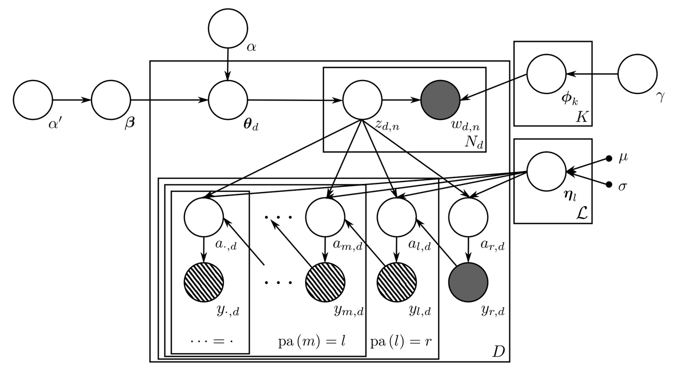


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [12], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{ -\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2} \right\} \quad (1)$$

Here, $c^{k,-(n,d)}_{v,d'}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}} \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l \right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right) \right\} \mathbb{I}(a_{l,d} y_{l,d} > 0), \quad (3)$$

is also a standard probit regression result [**?**].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [21]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [20]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{d,(\cdot)}, m\right) (\alpha\beta_k)^m \quad (5)$$

$s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [5].

Supervised latent Dirichlet allocation (sLDA) builds on LDA by incorporating supervision in the form of an observed exponential family response variable per document. As a result, sLDA infers topics such that the model predicts the response variable while improving word likelihood. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [4].

Other models that incorporate LDA and supervision include LabeledLDA[17], DiscLDA[13], and other models applied to computer vision and document networks[22, 6]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents in a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[**? ? ? ?**].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [8], and sometimes make mistakes [10].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [7, 11, 9]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [14, 18, 16, 19, 15].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation

The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

HSLDA	sLDA with Independent Regressors
children, animation, animated, kids, songs, story, family, new, fun,video, friends, music, live, classic, adventure, minute, world, magic,animals, song, musical, voiced, feature, along, baby, ages, magical, action, favorite, special, tale, little, child, holiday, help, voice, toys, original, learn, characters, computer, sing, like, adults, parents, version, old, first, including, young,	children, kids, animated, animation, songs, video, friends, fun, animals, family, story, new, little, music, like, minute, along, musical, song, adventure, help, classic, baby, child, ages, favorite, sing, learn, holiday, magical, tale, world, special, parents, live, way, day, boy, make, old, magic, adventures, home, show, gang, animal, adults, toys, many, action
comedy, funny, comic, hilarious, series, gags, jokes, slapstick, satire, laugh, stars, films, comedies, first, comedian, funniest, show, involving, characters, cast, perfect, mob, performance, play, writer, television, satirical, lines, hilariously, boss, routine, silent, funnier, comedic, chemistry, comedians, talk, routines, brothers, girlfriend, shorts, mouse, whose, cinema, screwball, hit, humor, brilliant, gag, new,	comedy, show, play, characters, sketches, funny, mob, performance, people, screwball, cast, best, including, stars, sketch, girlfriend, hilariously, charming, ensemble, comic, character, star, writer, hit, stage, host, classic, romantic, mutants, boss, new, played, together, choice, like, karaoke, daughter, vehicle, news, splitting, actress, get, story, hilarious, wacky, featuring, could, side, comedies, comedic
thriller, horror, murder, crime, killer, police, mystery, detective, cop, one, dead, suspense, plot, case, blood, drug, dark, mysterious, death, revenge, violence, town, gang, turns, prison, noir, violent, deadly, stars, young, woman, nbsp, body, vampire, night, murders, kill, criminal, action, becomes, wife, cops, gangster, sex, victims, victim, brutal, terror, evil, murdered	horror, film, vampire, blood, one, dead, supernatural, killer, evil, thriller, gore, monster, young, night, terror, mysterious, director, movie, films, cult, creepy, terrifying, original, dark, chilling, genre, atmosphere, victims, classic, murder, haunted, town, mystery, story, body, ghost, tale, family, murders, castle, house, nightmare, budget, eerie, death, doctor, victim, later, bloody, effects
workout, body, yoga, video, moves, dance, minute, program, fitness, easy, poses, muscles, routine, exercises, learn, strength, movements, practice, help, step, w, workouts, minutes, techniques, flexibility, get, exercise, one, instructor, first, series, time, instruction, use, follow, training, breathing, two, basic, back, ll, muscle, fun, warm, steps, technique, balance, beginners, designed, using	workout, body, yoga, video, moves, dance, minute, program, fitness easy, poses, muscles, exercises, learn, routine, strength, movements, minutes, get, help, practice, step, series, techniques, workouts, time, flexibility, exercise, instructor, instruction, one, two, fun, training, use, first, back, follow, steps, breathing, muscle, basic, using, ll, beginners, three, warm, balance, technique, designed

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[4] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[6] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

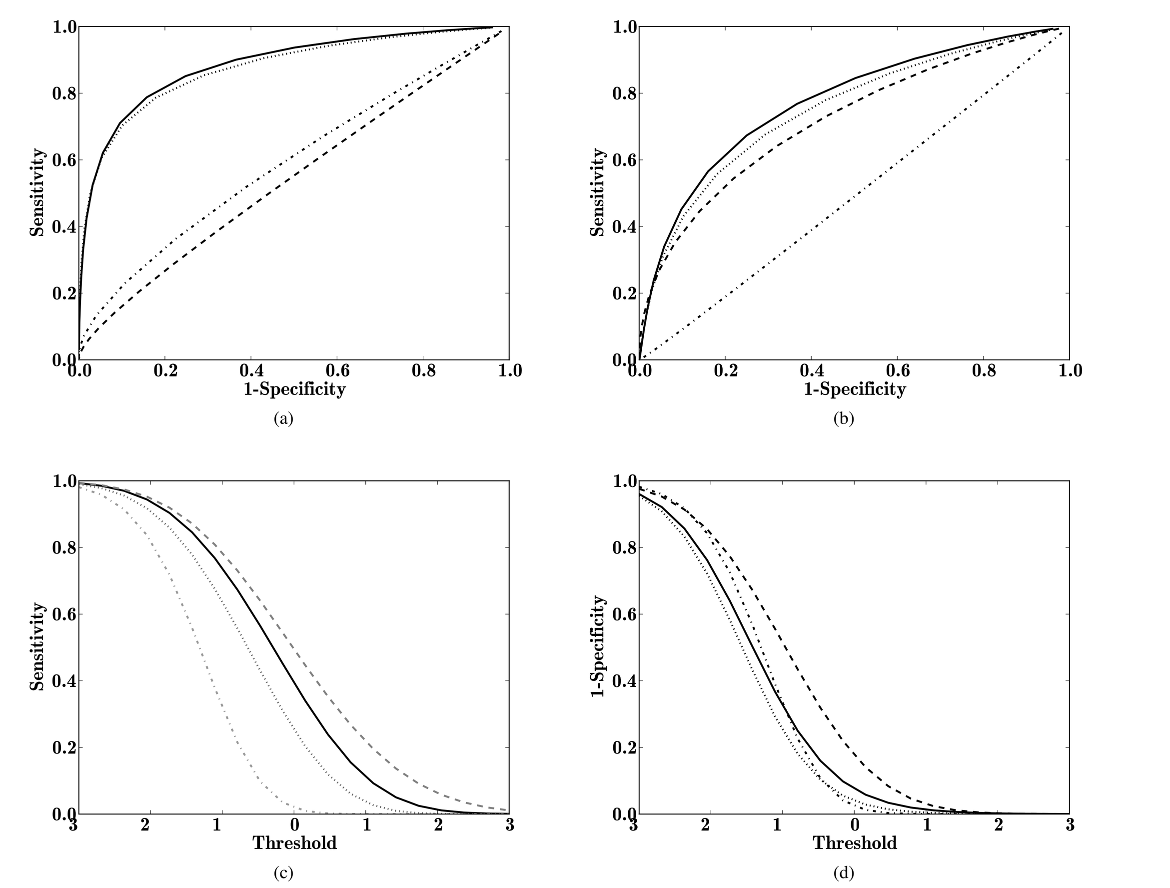


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

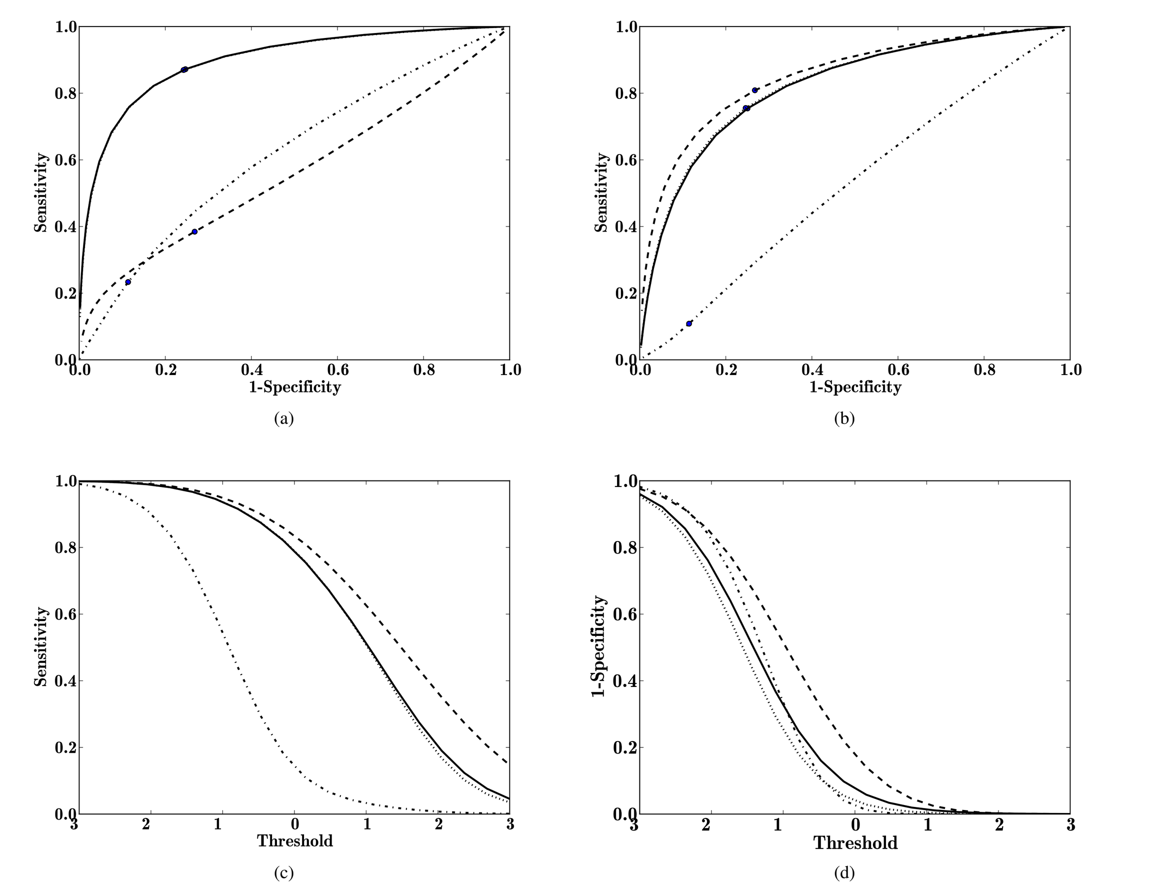


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[7] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[8] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[9] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[10] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[11] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[12] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[13] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[14] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[15] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[16] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[17] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[18] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[19] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[22] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include web pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurrence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to the health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

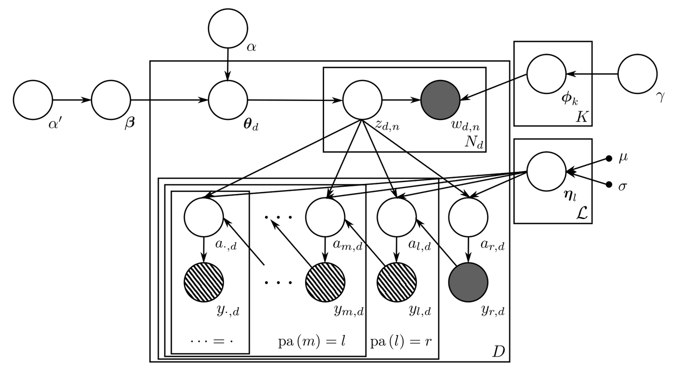


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu\mathbf{1}_K, \sigma\mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha'\mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol\theta_d \mid \boldsymbol\beta, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol\beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol\theta_d \sim \mathrm{Multinomial}(\boldsymbol\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol\phi_{1:K} \sim \mathrm{Multinomial}(\boldsymbol\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol\eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1}\sum_{n=1}^{N_d}\mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [?]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol\eta_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [13], we have analytically marginalized out the parameters $\boldsymbol\phi_{1:K}$ and $\boldsymbol\theta_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol\eta$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol\eta, \alpha, \boldsymbol\beta, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\boldsymbol\beta_k\right)\frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)}\prod_{l \in \mathcal{L}_d}\exp\left\{-\frac{(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l - a_{l,d})^2}{2}\right\} \quad (1)$$

Here, $c^{k,-(n,d')}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [?]; the specific form of the update is a standard result from Bayesian normal linear regression [?]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol\eta) \propto \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol\eta_l^T\bar{\mathbf{z}}_d\right)\right\}\mathbb{I}(a_{l,d}y_{l,d} > 0). \quad (3)$$

is also a standard probit regression result [4].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol\beta$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [22]. Sampling $\boldsymbol\beta$ is done using the "direct assignment" method of Teh et al. [21]

$$\boldsymbol\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol\beta$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol\beta\right) = \frac{\Gamma(\alpha\boldsymbol\beta_k)}{\Gamma\left(\alpha\boldsymbol\beta_k + c^k_{\cdot,d}\right)}s\left(c^k_{\cdot,d}, m\right)(\alpha\boldsymbol\beta_k)^m \quad (5)$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include rating associated with an online eview, grades for an essay, and number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[18], DiscLDA[14], and other models applied to computer vision and document networks[23, 7]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents in a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[**? ? ? ?**].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [9], and sometimes make mistakes [11].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [8, 12, 10]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [15, 19, 17, 20, 16].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We applied HSLDA, along with three other closely related models, to the clinical data and the retail product data. Specifically, we evaluate models including sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

We evaluated model performance for all three models with a range of values for μ ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$), the mean prior parameter for regression parameters.

5.3 Evaluation

The two main methods of evaluation for the model are prediction and topic quality. To evaluate predictive performance for all comparison models equivalently, each model was evaluated with two methods. The first evaluation method augments the observed labels in the held out set with their ancestors and considers all other non-existant labels to be negative. The second method ignores the ancestors of the observed labels in the held out set and considers all other non-existant labels to be negative. This uniform treatment of ancestors allows for a fair comparison of the models.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

The two measures for predictive performance used here include the true positive rate and the false positive rate evaluated based on $p\left(y_{l,\hat{d}} \mid w_{1:N,d}\right)$ for each label in each model.

6 Results

HSLDA	sLDA with Independent Regressors
children, animation, animated, kids, songs, story, family, new, fun,video, friends, music, live, classic, adventure, minute, world, magic,animals, song, musical, voiced, feature, along, baby, ages, magical, action, favorite, special, tale, little, child, holiday, help, voice, toys, original, learn, characters, computer, sing, like, adults, parents, version, old, first, including, young,	children, kids, animated, animation, songs, video, friends, fun, animals, family, story, new, little, music, like, minute, along, musical, song, adventure, help, classic, baby, child, ages, favorite, sing, learn, holiday, magical, tale, world, special, parents, live, way, day, boy, make, old, magic, adventures, home, show, gang, animal, adults, toys many, action
comedy, funny, comic, hilarious, series, gags, jokes, slapstick, satire, laugh, stars, films, comedies, first, comedian, funniest, show, involving, characters, cast, perfect, mob, performance, play, writer, television, satirical, lines, hilariously, boss, routine, silent, funnier, comedic, chemistry, comedians, talk, routines, brothers, girlfriend, shorts, mouse, whose, cinema, screwball, hit, humor, brilliant, gag, new,	comedy, show, play, characters, sketches, funny, mob, performance, people, screwball, cast, best, including, stars, sketch, girlfriend, hilariously, charming, ensemble, comic, character, star, writer, hit, stage, host, classic, romantic, mutants, boss, new, played, together, choice, like, karaoke, daughter, vehicle, news, splitting, actress, get, story, hilarious, wacky, featuring, could, side, comedies, comedi
thriller, horror, murder, crime, killer, police, mystery, detective, cop, one, dead, suspense, plot, case, blood, drug, dark, mysterious, death, revenge, violence, town, gang, turns, prison, noir, violent, deadly, stars, young, woman, nbsp, body, vampire, night, murders, kill, criminal, action, becomes, wife, cops, gangster, sex, victims, victim, brutal, terror, evil, murdered	horror, film, vampire, blood, one, dead, supernatural, killer, evil, thriller, gore, monster, young, night, terror, mysterious, director, movie, films, cult, creepy, terrifying, original, dark, chilling, genre, atmosphere, victims, classic, murder, haunted, town, mystery story, body, ghost, tale, family, murders, castle, house, nightmare, budget, eerie, death, doctor, victim, later, bloody, effects
workout, body, yoga, video, moves, dance, minute, program, fitness, easy, poses, muscles, routine, exercises, learn, strength, movements, practice, help, step, w, workouts, minutes, techniques, flexibility, get, exercise, one, instructor, first, series, time, instruction, use, follow, training, breathing, two, basic, back, ll, muscle, fun, warm, steps, technique, balance, beginners, designed, using	workout, body, yoga, video, moves, dance, minute, program, fitness easy, poses, muscles, exercises, learn, routine, strength, movements, minutes, get, help, practice, step, series, techniques, workouts, time, flexibility, exercise, instructor, instruction, one, two, fun, training, use, first, back, follow, steps, breathing, muscle, basic, using, ll, beginners, three, warm, balance, technique, designed

7 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[7] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[8] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

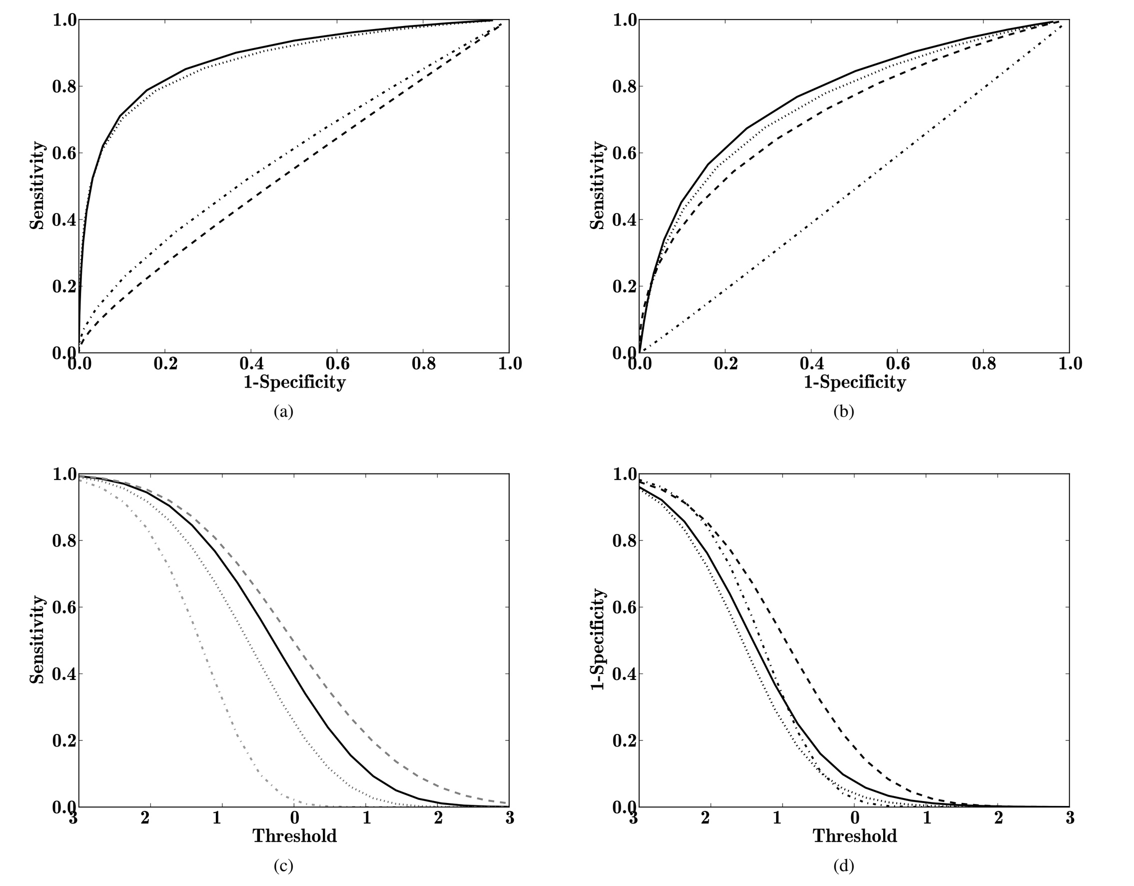


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

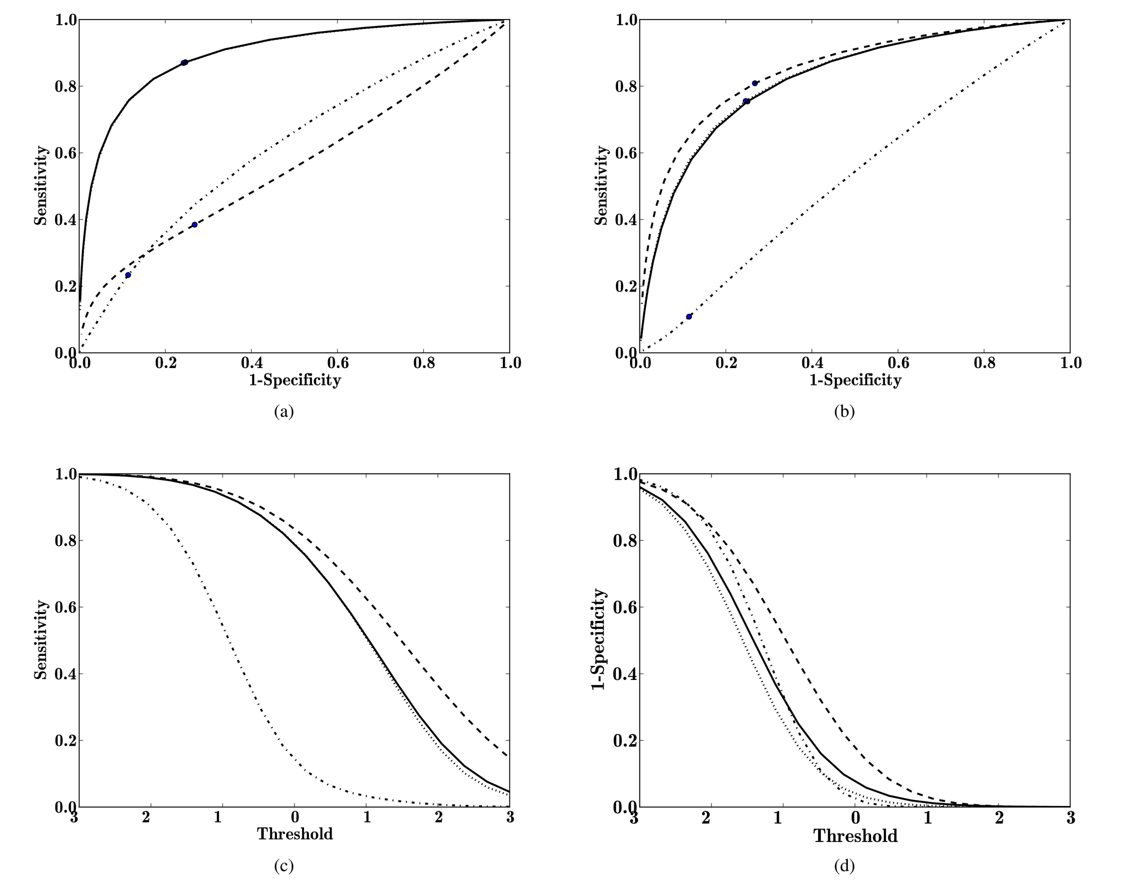


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[9] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[10] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[11] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[12] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[13] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[14] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[15] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[16] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[17] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[19] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[20] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[22] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[23] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include web pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurrence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

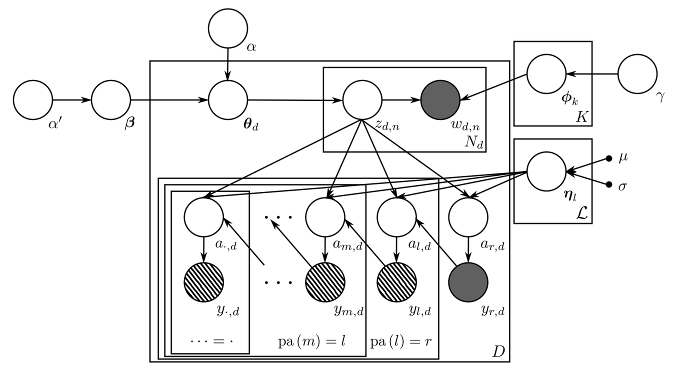


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**?** ?]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [15], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d')}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

is also a standard probit regression result [4].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{\cdot,d}^k\right)} s\left(c_{\cdot,d}^k, m\right)\left(\alpha \boldsymbol{\beta}_k\right)^m \tag{5}$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include rating associated with an online eview, grades for an essay, and number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[22], DiscLDA[17], and other models applied to computer vision and document networks[27, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents in a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[20, 10, 16, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 14, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [18, 23, 21, 24, 19].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression parameters ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$). The number of topics for all models was set to 50, the parameters for the prior distributions α, α^prime, and γ were a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) these potentially erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[15] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[16] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[17] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

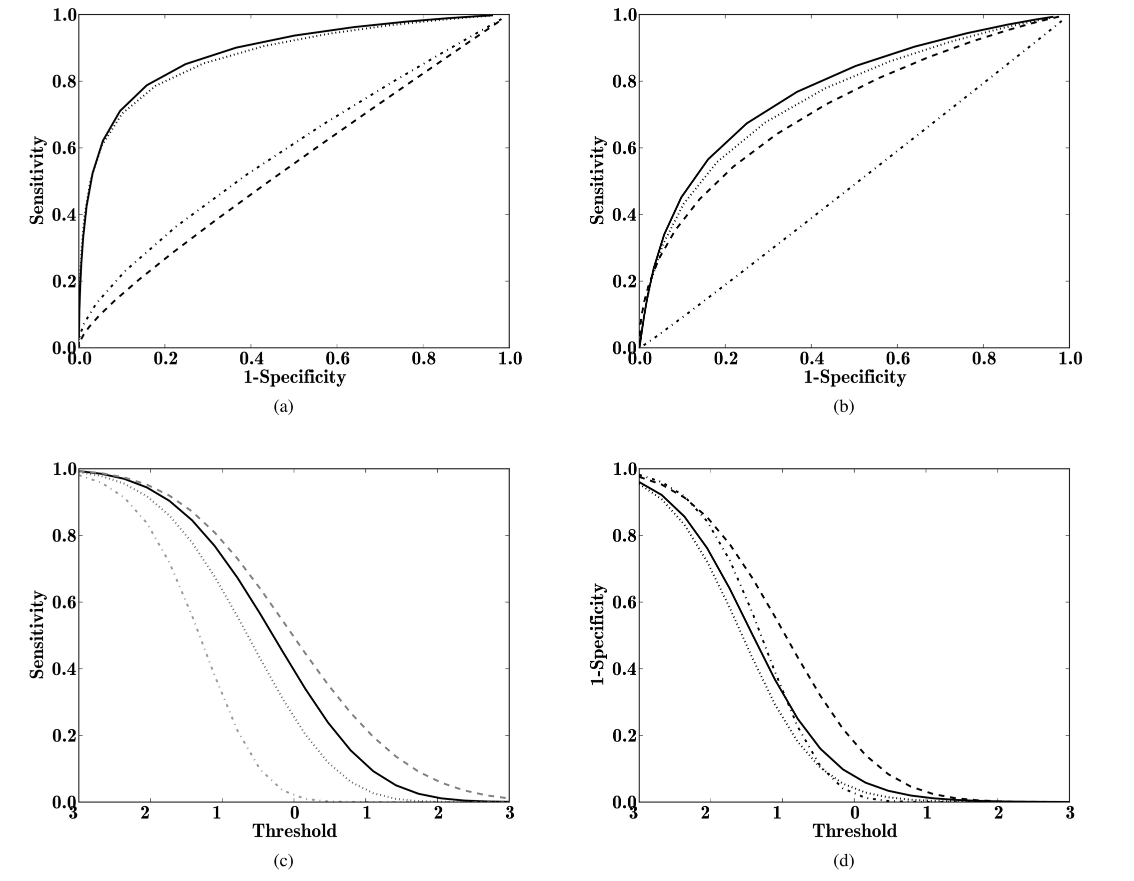


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

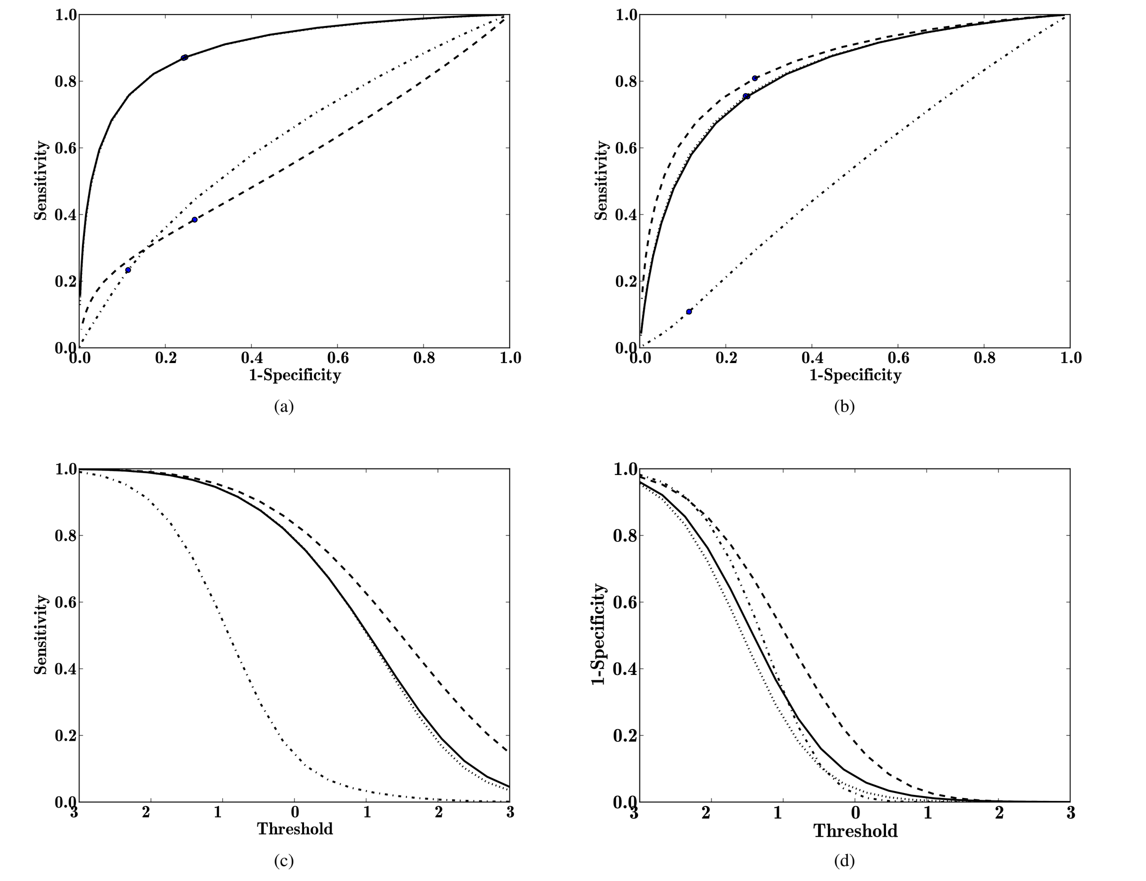


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[18] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[19] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[20] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[21] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[23] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[24] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

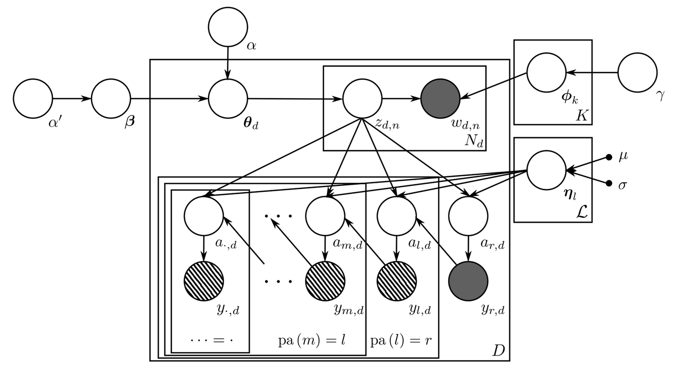


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$ both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to decide whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [15], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \quad (1)$$

Here, $c_{v,d'}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d'. The $(\cdot)$ in the subscript means the count resulting from summing over the omitted subscript variable. Also $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [**?**]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. That the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

is also a standard probit regression result [4].

HSLDA departs from stock LDA in that we estimate a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following distribution.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{d,(\cdot)}^k\right)} s\left(c_{d,(\cdot)}^k, m\right) (\alpha\beta_k)^m \quad (5)$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include rating associated with an online review, grades for an essay, and number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[22], DiscLDA[17], and other models applied to computer vision and document networks[27, 8]. These models, however, do not implement constraints on the label space.

In other work, researchers have classified documents in a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[20, 10, 16, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 14, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [18, 23, 21, 24, 19].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression parameters ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$). The number of topics for all models was set to 50, the parameters for the prior distributions α, $\alpha^{p}rime$, and γ were a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) these potentially erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[15] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[16] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[17] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

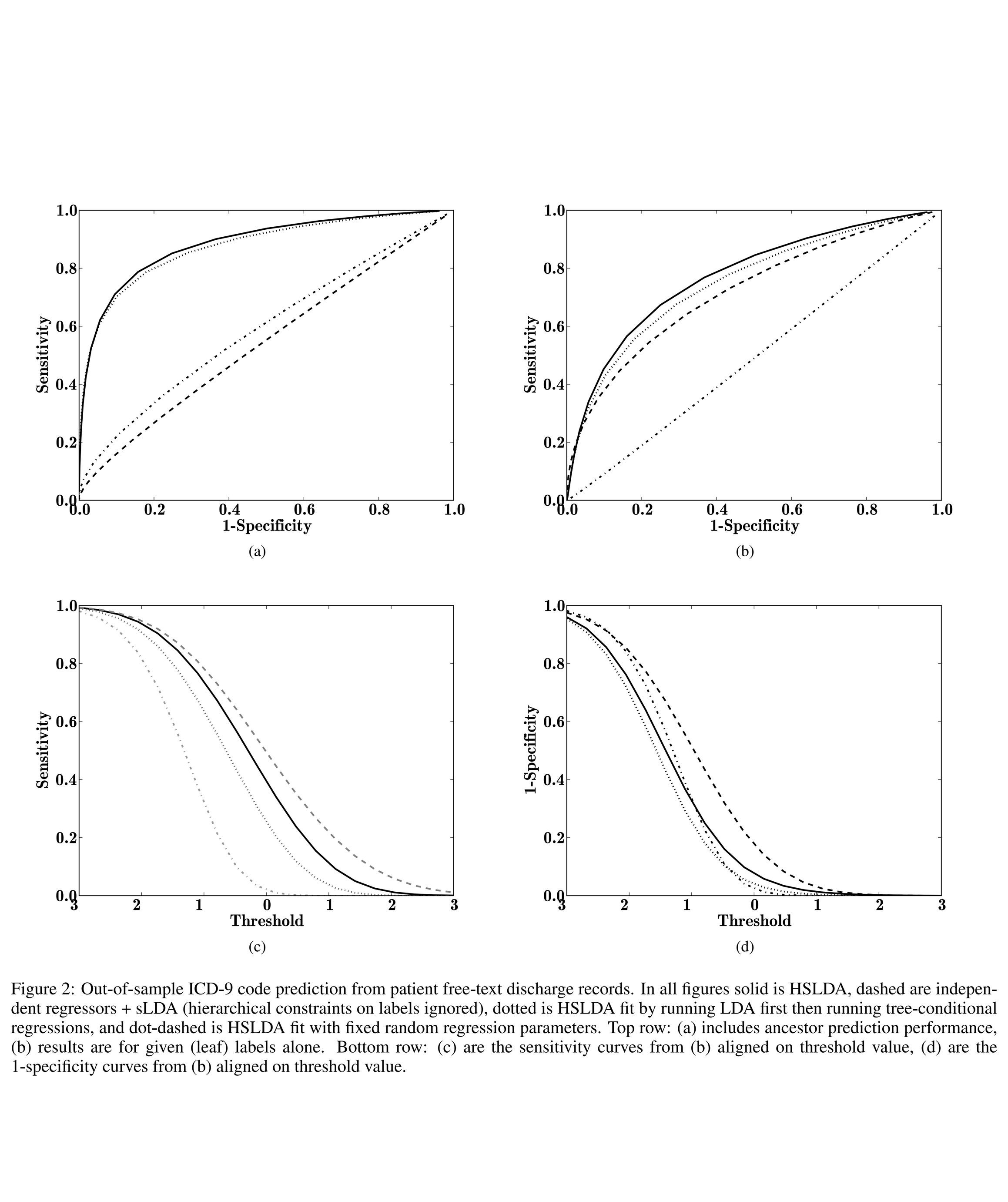


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

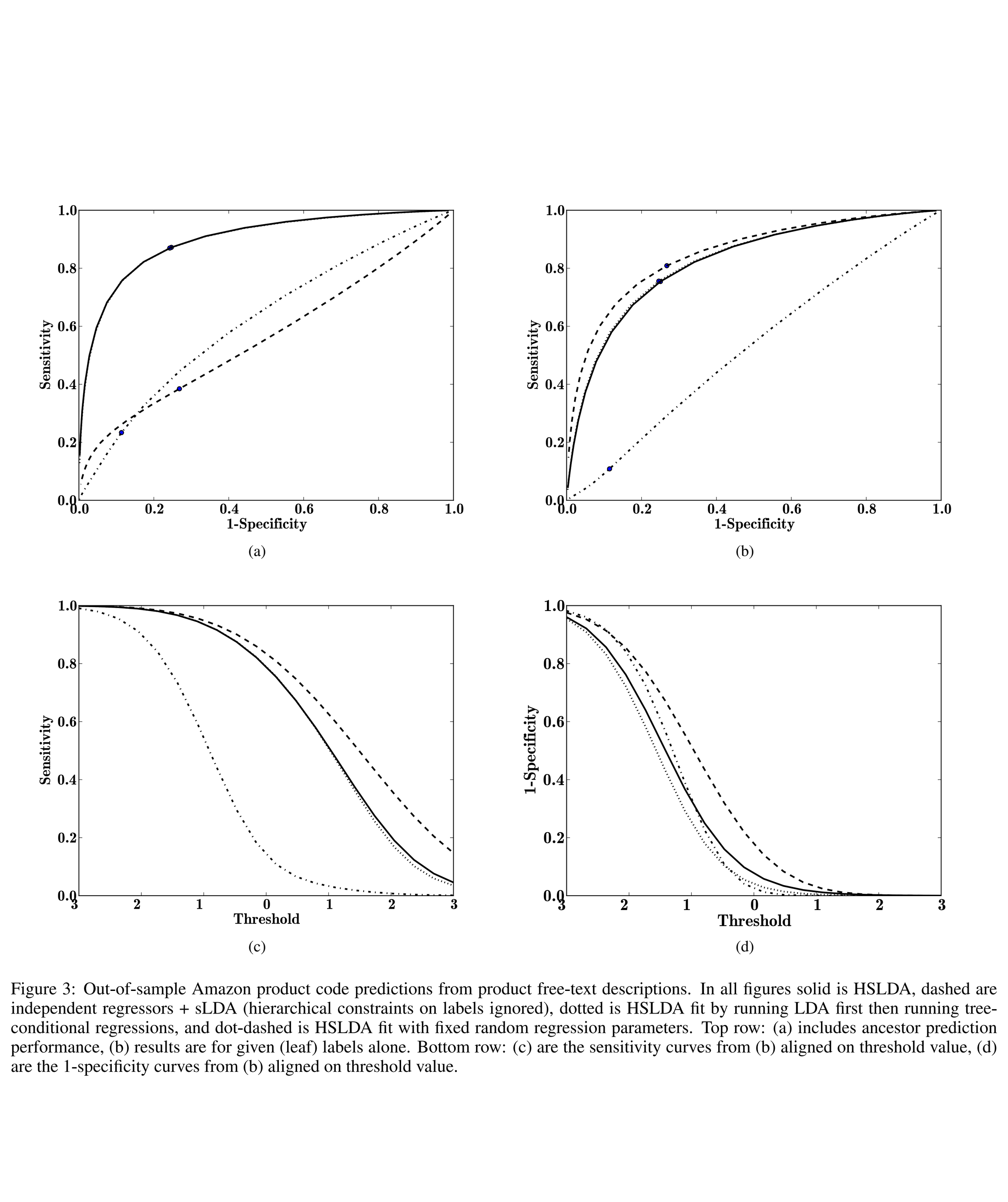


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[18] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[19] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[20] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[21] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[23] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[24] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include page and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

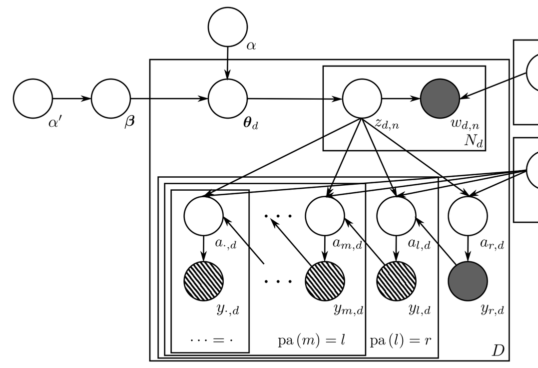


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [15], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following, $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The $(\cdot)$ in the subscript indicates that the related count is a sum over the omitted subscript variable. Also, $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [**?**]. It is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4].

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA implements an asymmetric prior as a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{\cdot,d}^k\right)} s\left(c_{\cdot,d}^k, m\right) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[22], DiscLDA[17], and other models applied to computer vision and document networks[27, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[20, 10, 16, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 14, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [18, 23, 21, 24, 19].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression parameters ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$). The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) these potentially erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[15] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[16] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[17] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

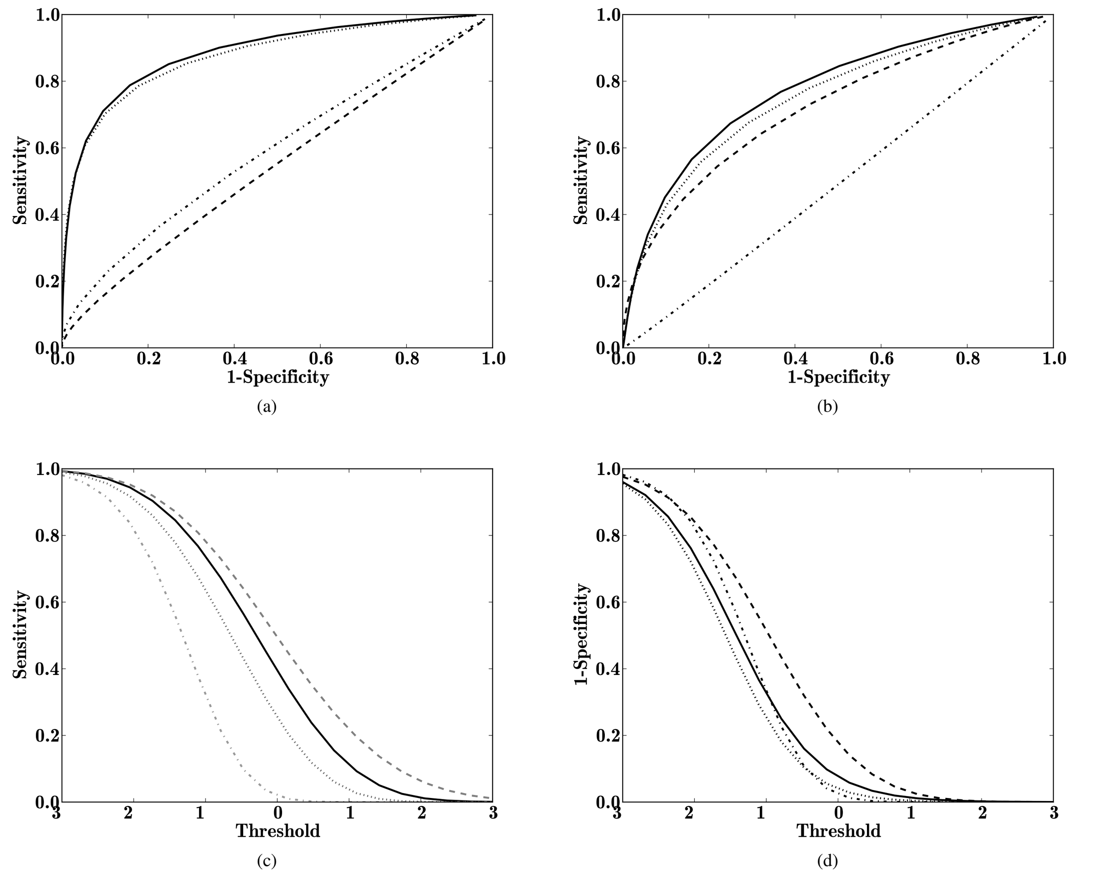


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

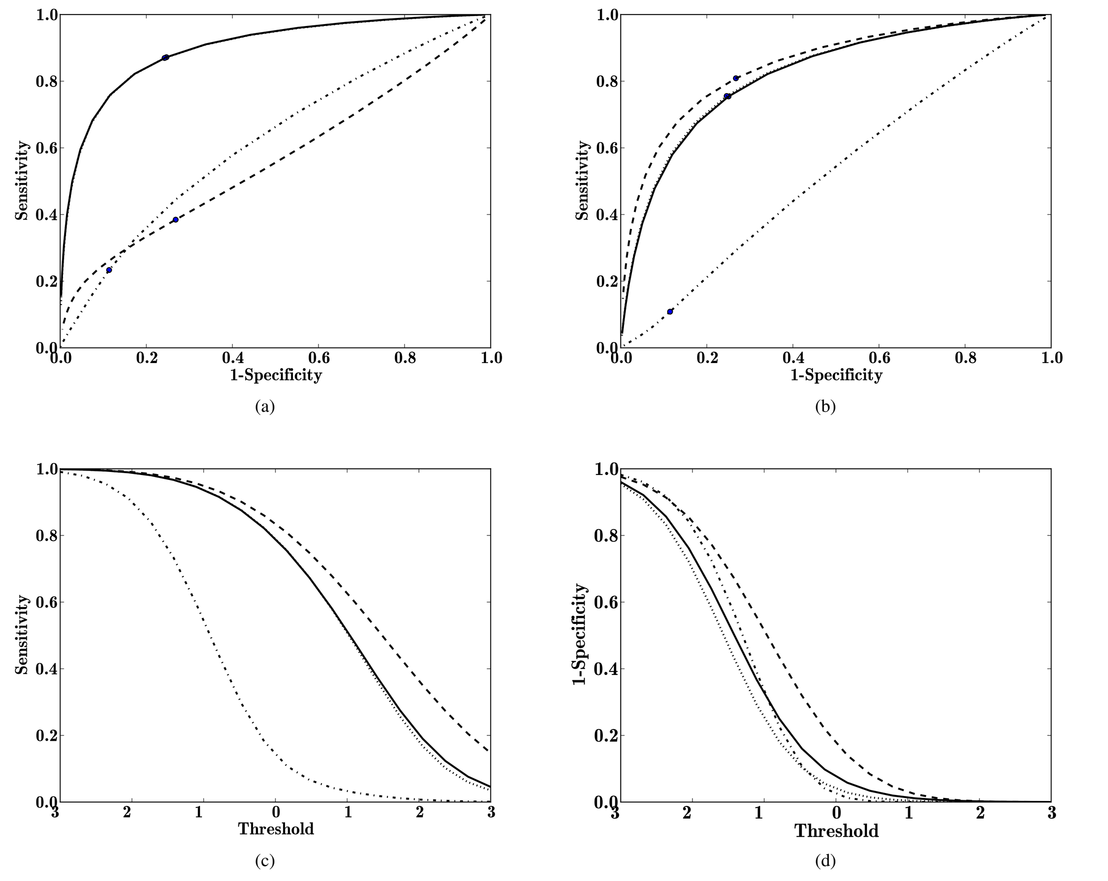


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[18] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[19] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[20] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[21] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[23] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[24] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

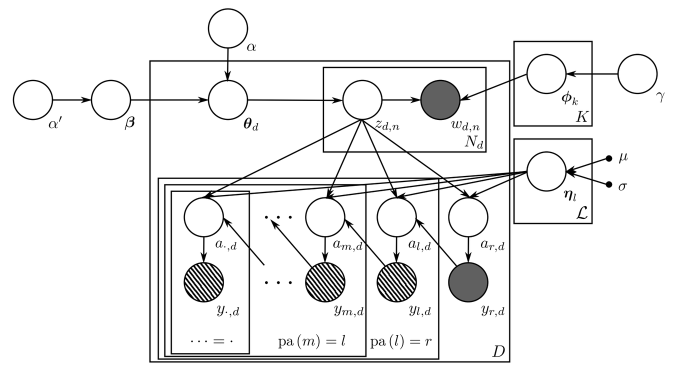


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [15], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following, $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \quad (1)$$

Here, $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The $(\cdot)$ in the subscript indicates that the related count is a sum over the omitted subscript variable. Also, $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [?]. It is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA implements an asymmetric prior as a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{d,\cdot}^k\right)} s\left(c_{d,(\cdot)}^k, m\right)(\alpha\beta_k)^m \quad (5)$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[22], DiscLDA[17], and other models applied to computer vision and document networks[27, 8]. These models, however, do not implement constraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[20, 10, 16, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 14, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [18, 23, 21, 24, 19].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) these potentially erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[14] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[15] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

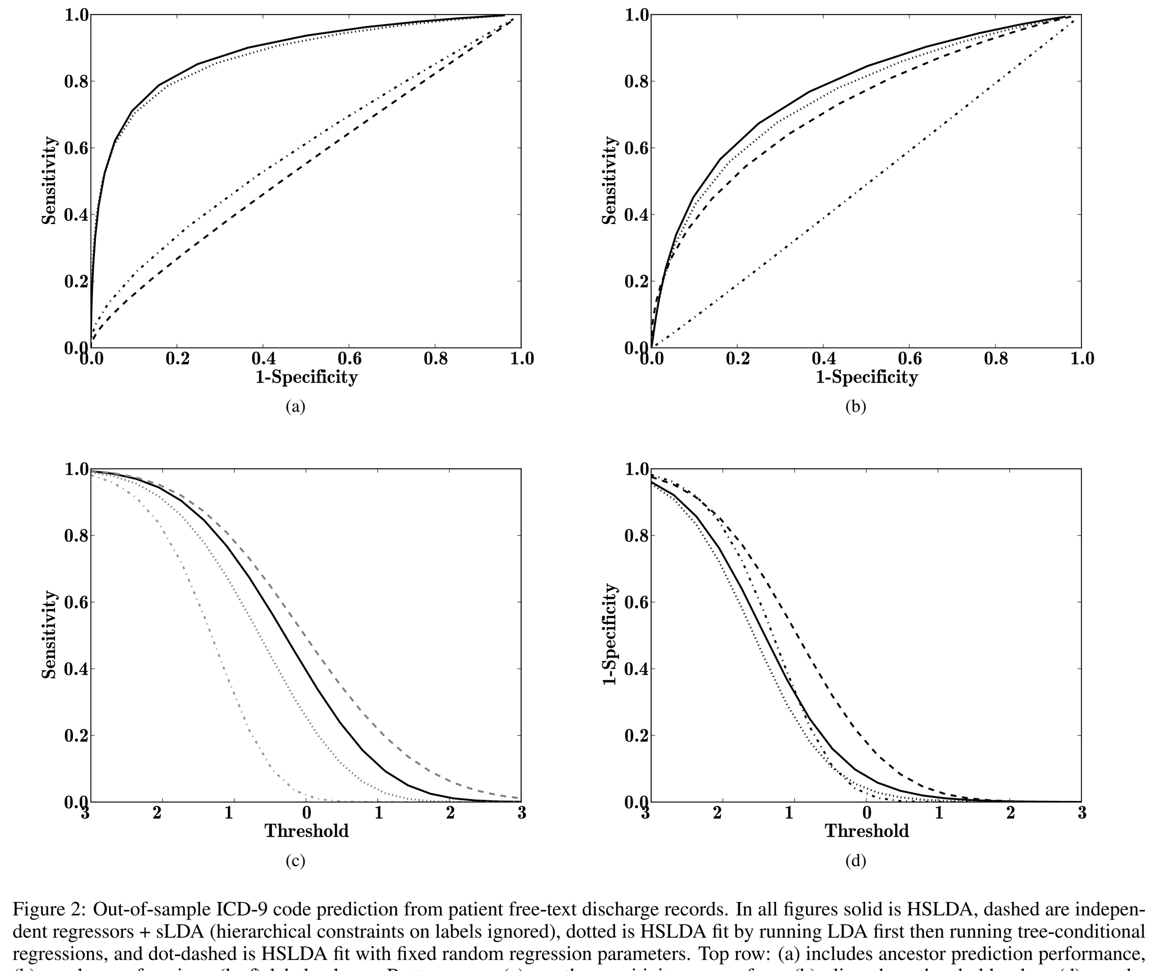


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

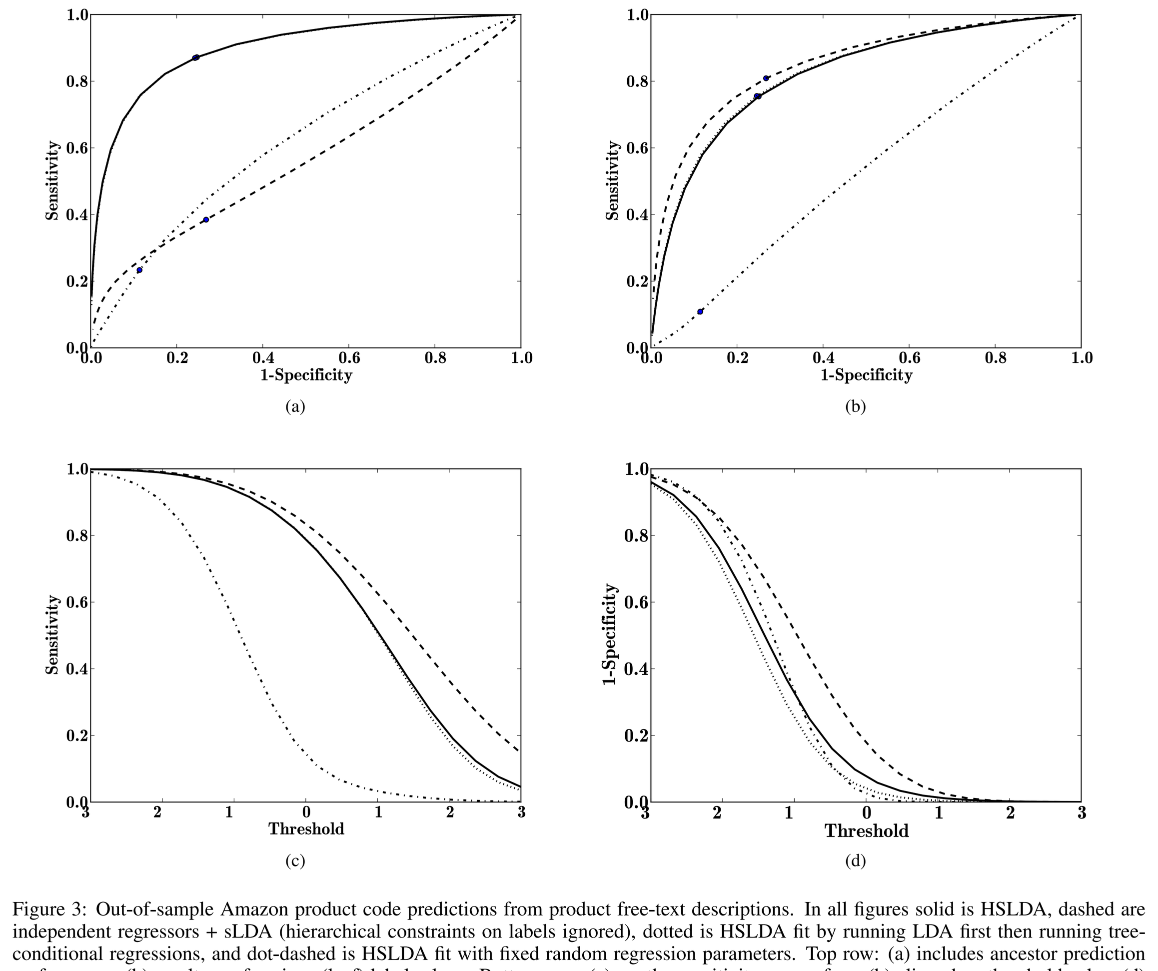


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[16] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[17] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[18] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[19] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[20] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[21] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[23] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[24] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include pages and their placement in Web directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to structure-agnostic models. Furthermore, we show evidence that the resulting HSLDA topics are more descriptive of the underlying data than sLDA topics, which ignore the label hierarchy.

1 Introduction

The task of multi-label classification, selecting the k-best labels for a given instance, has been a topic of research for several years. One simplistic way to carry out the classification is through a series of independent binary classifiers, but this ignores the many inherent dependencies among the labels. Thus, much work has been devoted on incorporating the co-occurence patterns of the labels into the classification task. In this paper, we focus on multi-label classification, where the labels are organized in a hierarchical structure. Scenarios of use include, but are not limited to, placing webpages into manually curated Internet directories [2], categorizing images according to a taxonomy, tagging product descriptions with catalogue information [3], and assigning diagnosis codes to clinical records [1].

There are several challenges entailed in incorporating the hierarchical nature of labels into the classification task. One pertains to the labeling itself: in the datasets (especially real-world, noisy ones), for a given label, instances labeled with it contribute positive instance, but it is unclear how to determine the negative instances. In particular, how to treat the parent labels of the selected ones?

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [5] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering labels as a flat list.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Following the trend observed in supervised topic modeling, we note that the learned topic models are more representative of the underlying data in both of our datasets [5].

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

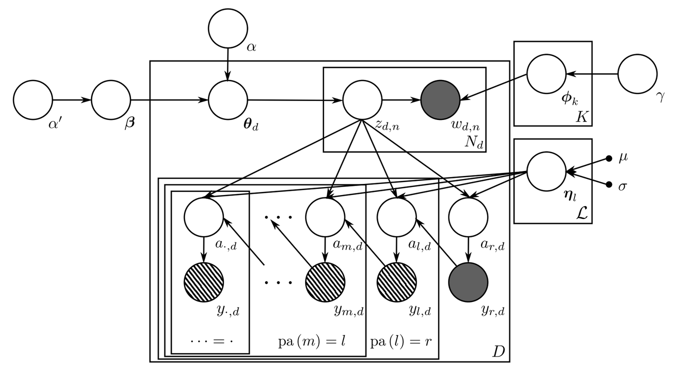


Figure 1: HSLDA graphical model

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ

parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following, $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

Here, $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The $(\cdot)$ in the subscript indicates that the related count is a sum over the omitted subscript variable. Also, $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4].

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA implements an asymmetric prior as a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \tag{4}$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + c_{\cdot,d}^k\right)} s\left(c_{\cdot,d}^k, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \tag{5}$$

$s\left(n, m\right)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[23], DiscLDA[18], and other models applied to computer vision and document networks[28, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[21, 10, 17, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 25, 20].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p\left(\alpha\right)$, $p\left(\alpha'\right)$, and $p\left(\gamma\right)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

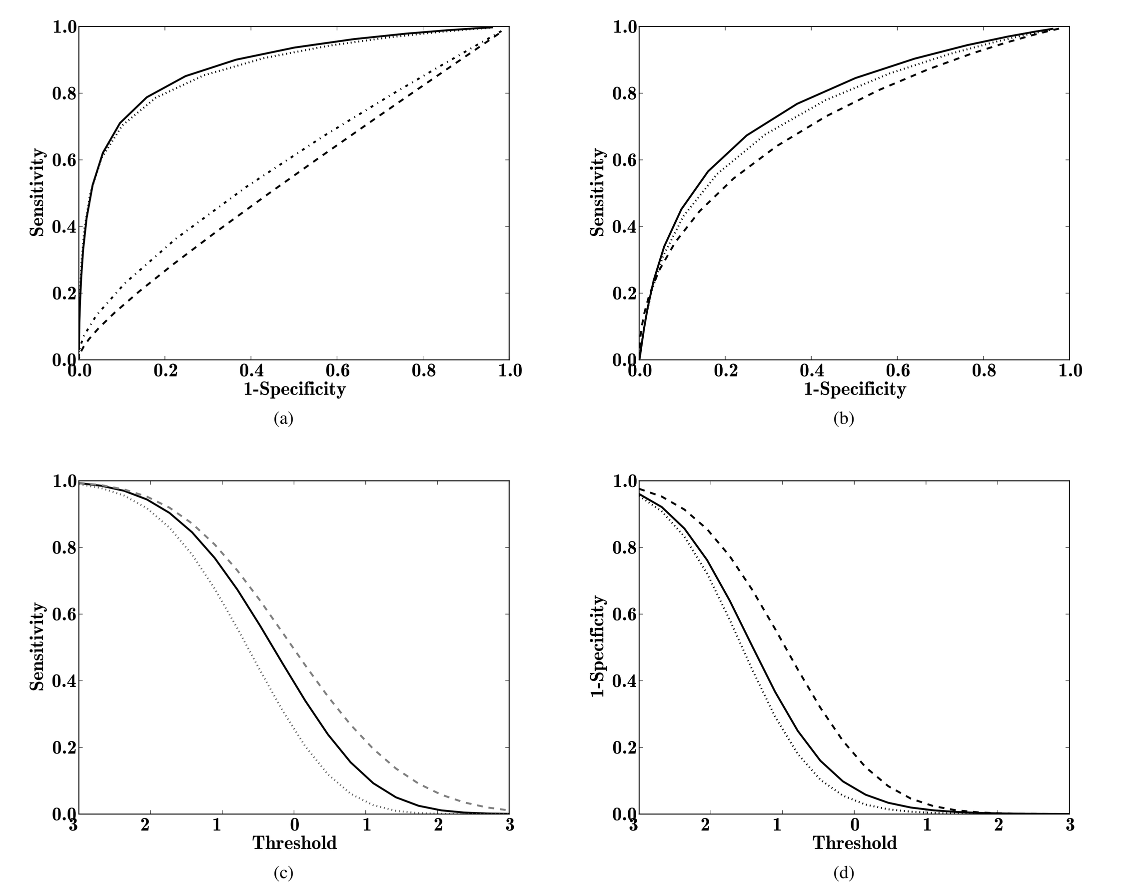


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

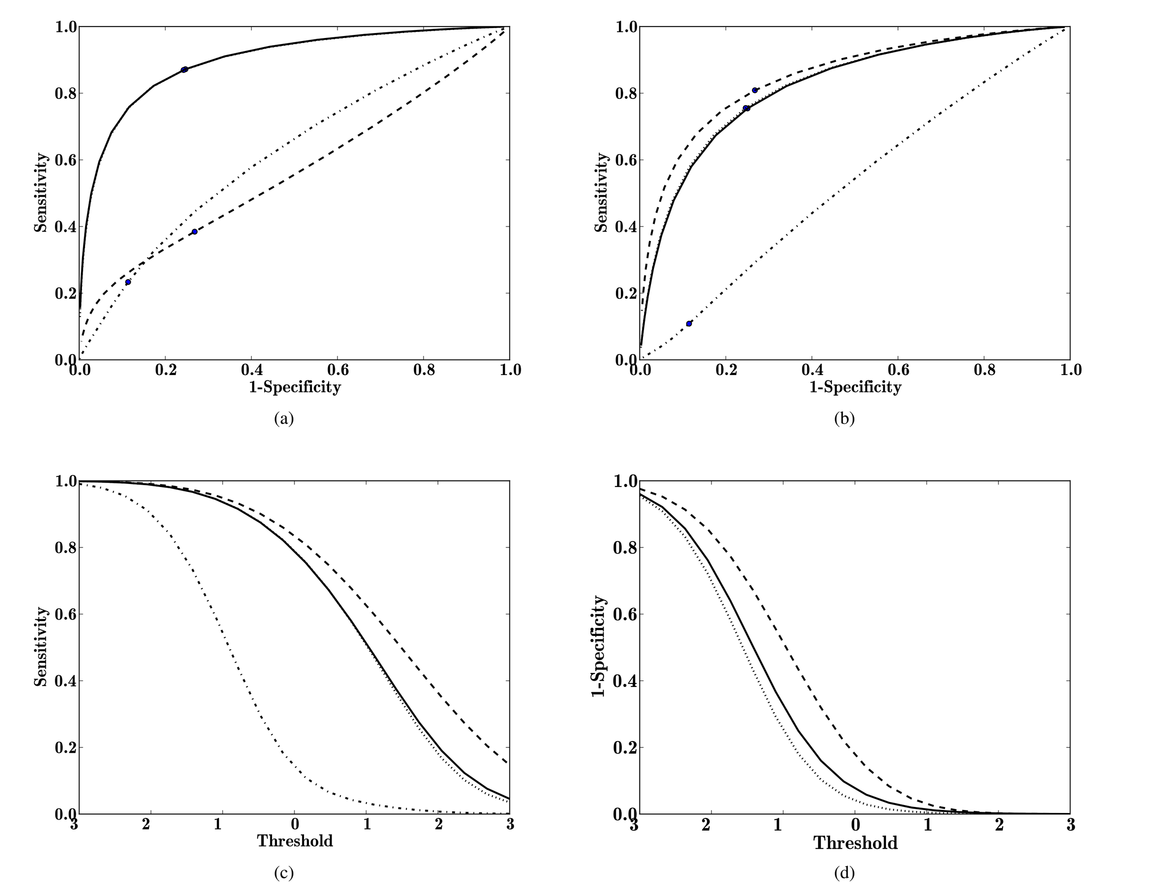


Figure 3: Out-of-sample Amazon product code predictions from product free-text descriptions. In all figures solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), dotted is HSLDA fit by running LDA first then running tree-conditional regressions, and dot-dashed is HSLDA fit with fixed random regression parameters. Top row: (a) includes ancestor prediction performance, (b) results are for given (leaf) labels alone. Bottom row: (c) are the sensitivity curves from (b) aligned on threshold value, (d) are the 1-specificity curves from (b) aligned on threshold value.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[21] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[22] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [**?**] as available from [3]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [**?**]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label l except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

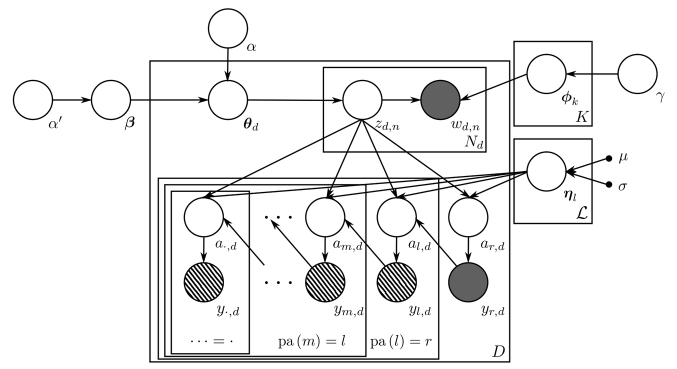


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**?** **?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. In the following, $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, $\boldsymbol{\eta}$ is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \quad (1)$$

Here, $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The $(\cdot)$ in the subscript indicates that the related count is a sum over the omitted subscript variable. Also, $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA implements an asymmetric prior as a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$ and are governed by the following function.

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

$s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[23], DiscLDA[18], and other models applied to computer vision and document networks[28, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[21, 10, 17, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review

the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 25, 20].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

(a)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (b) shows predictive perfmance as a function of the prior mean on regression parameters.

(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (b) shows predictive perfmance as a function of the prior mean on regression parameters.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[21] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[22] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [?] is available from [3]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [?]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

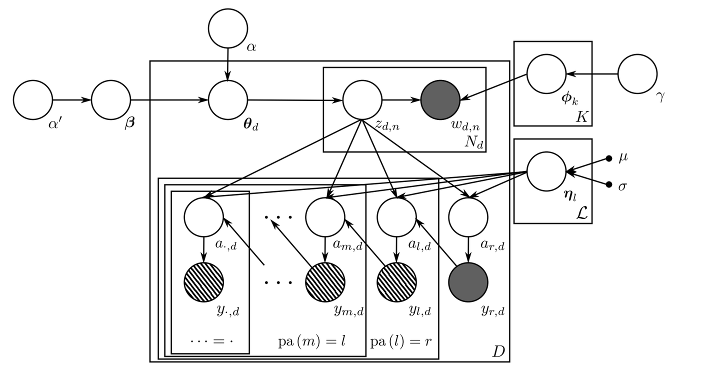


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$
     – Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     – Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     – Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T\eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     – Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise}\end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1}\sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to Gibbs sample the HSLDA posterior distribution. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$. In the following, $\mathbf{a}$ is the set of all auxiliary variables, $\mathbf{w}$ is the set of all words, η is the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ is the set $\mathbf{z}_d$ with element $z_{n,d}$ removed.

$$p(z_{n,d} = k \mid \mathbf{z}_d\backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d})^2}{2}\right\} \quad (1)$$

Here, $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The $(\cdot)$ in the subscript indicates that the related count is a sum over the omitted subscript variable. Also, $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA implements an asymmetric prior over topic assignments (i.e. β is a parameter in our model). This has been shown to improve the quality and stability of inferred topics [27]. Sampling β is done using the "direct assignment" method of Teh et al. [26]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

where $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β and are governed by the following function.

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right)(\alpha\beta_k)^m \quad (5)$$

$s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[23], DiscLDA[18], and other models applied to computer vision and document networks[28, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[21, 10, 17, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review

the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 25, 20].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

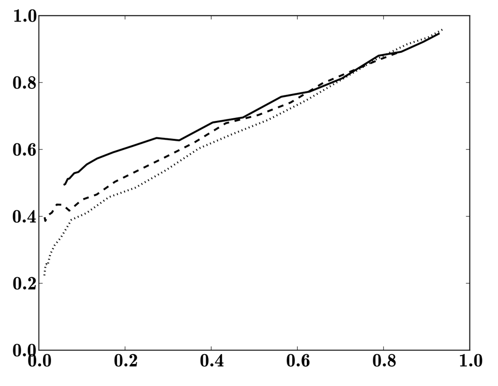
The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.



(a)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

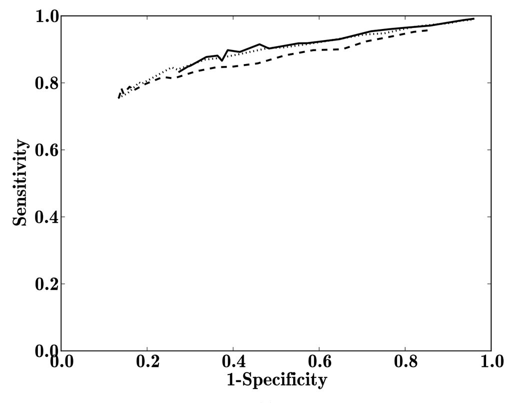
...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[21] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[22] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs (e.g. [?] as available from [3]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [?]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

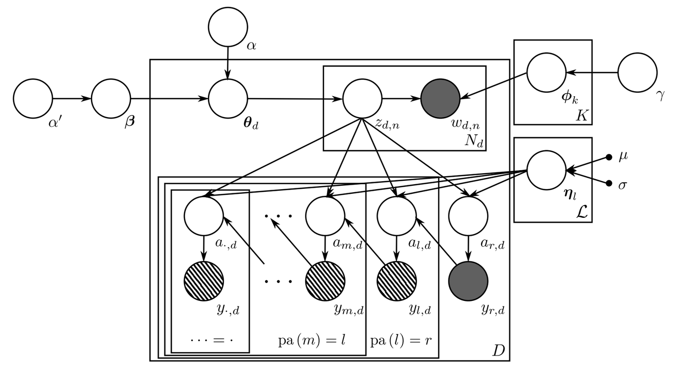


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[4]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \beta_k)}{\Gamma\left(\alpha \beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) (\alpha \beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

While there has been much work in multi-label classification of text and text modeling in general, we focus here on topic modeling approaches. Latent Dirichlet allocation (LDA) is a generative probabilistic model which represents documents as a mixed-membership bag of word. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [6]. sLDA is latent Dirichlet allocation (LDA) [6] augmented with per-document labeling, often taking the form of a single numerical or categorical label. Examples of labels include ratings associated with online reviews, grades for essays, and the number of times a webpage is linked. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [5].

Other models that incorporate LDA and supervision include LabeledLDA[23], DiscLDA[18], and other models applied to computer vision and document networks[28, 8]. These models, however, do not implement contraints on the label space.

In other work, researchers have classified documents into a hierarchy with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces and has focused on single label classification without a model of documents such as LDA[21, 10, 17, 7].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [11], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [9, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 25, 20].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [5]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

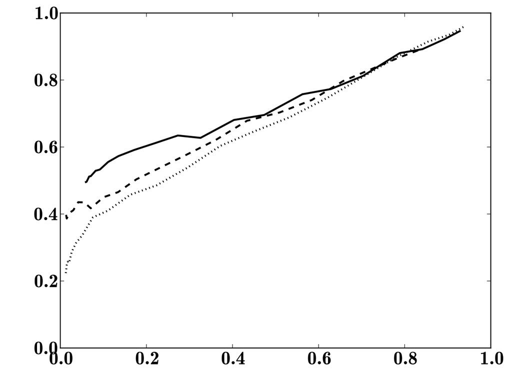
The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that HSLDA model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were considered positive and unobserved nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.



(a)

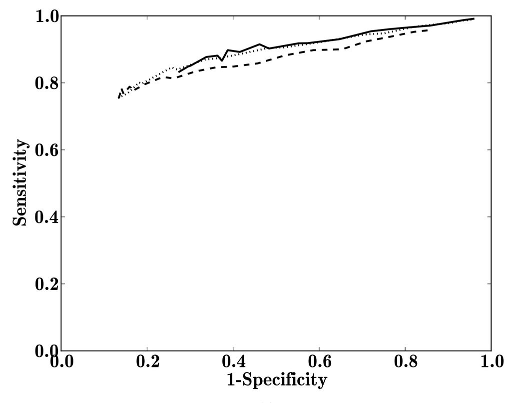
Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[]

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[9] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[10] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[11] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[12] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[21] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[22] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

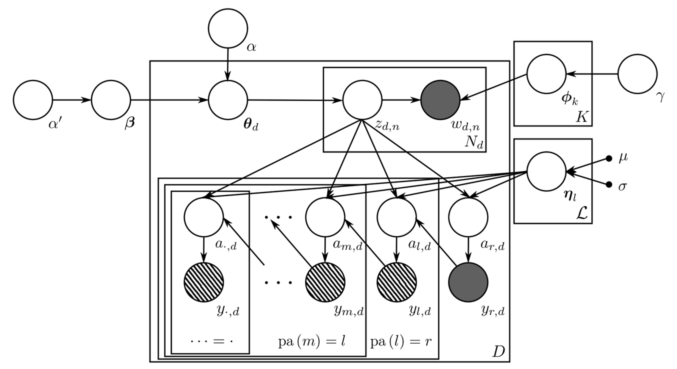


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha \beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [? ?]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label pa(l) is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, η the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \tag{2}$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. β is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling β is done using the "direct assignment" method of Teh et al. [28]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) (\alpha\beta_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 12, 19, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [13], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. This study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{\tilde{l},\tilde{d}} \mid w_{1:N_{\tilde{d}},\tilde{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\tilde{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\tilde{d}}$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

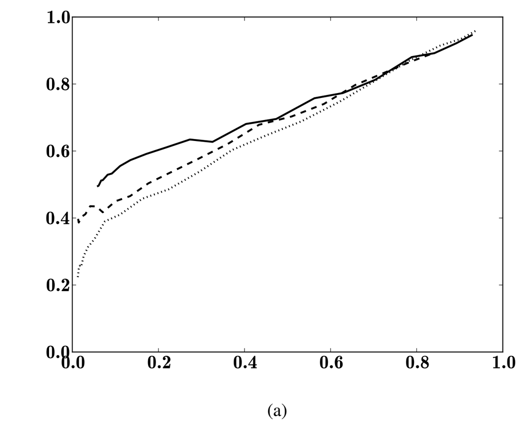


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

6 Discussion

We have described a mixed membership model with hierarchical supervision. We have demonstrated this model in the context of document modeling with hierarchical multi-label supervision. Such a model is appropriate in domains where there are hierarchical constraints among the labels such as is the case in an IS-A hierarchy.

...

- what about the nonparametric version of this?
- discuss the broader goal, from the beginning of search engine time, to combine categorization and free text. this, to our knowledge, is the first principled approach to doing so

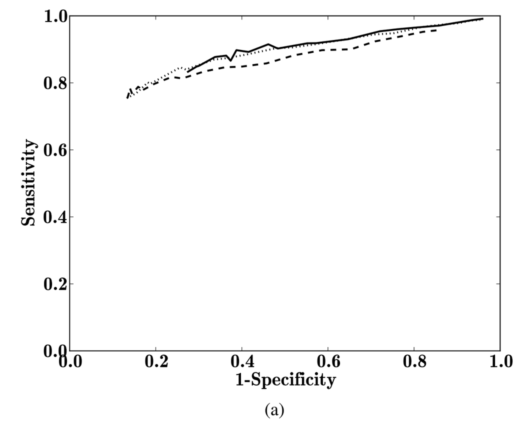


Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[1]

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[6] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[23] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[24] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

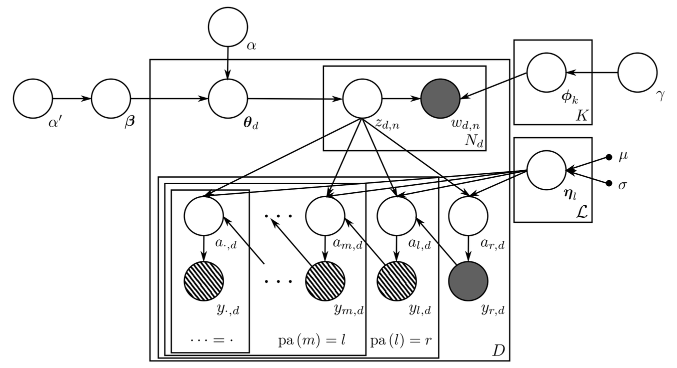


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \qquad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \qquad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \qquad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [28]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \qquad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha\beta_k)^m \qquad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 12, 19, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [13], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Logistic Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. This study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

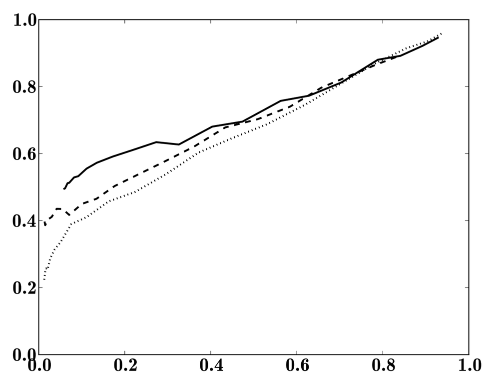
The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.



(a)

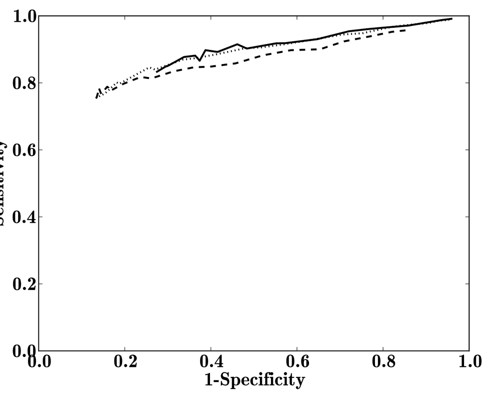
Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[1]

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[6] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[23] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[24] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

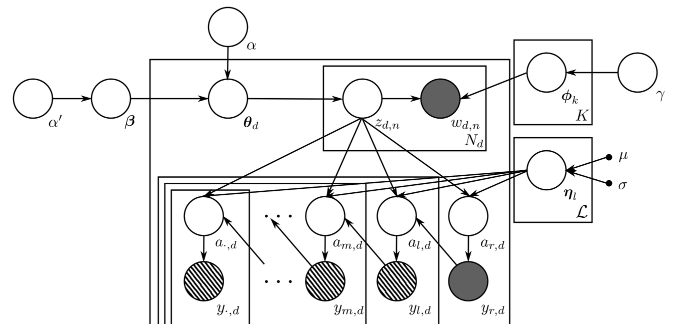


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, η the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d})^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. β is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling β is done using the "direct assignment" method of Teh et al. [28]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 12, 19, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [13], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by running LDA first followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied after inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

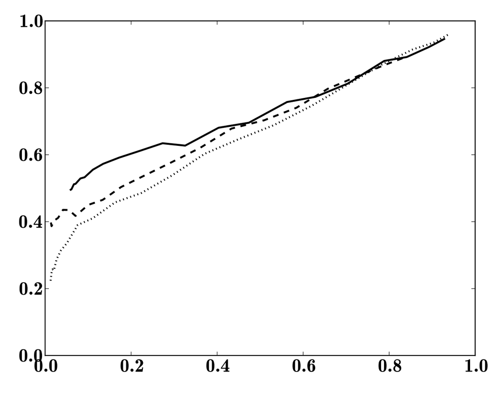
Figures 2 and 3 show the predictive performance of HSLDA relative to the three comparison models.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

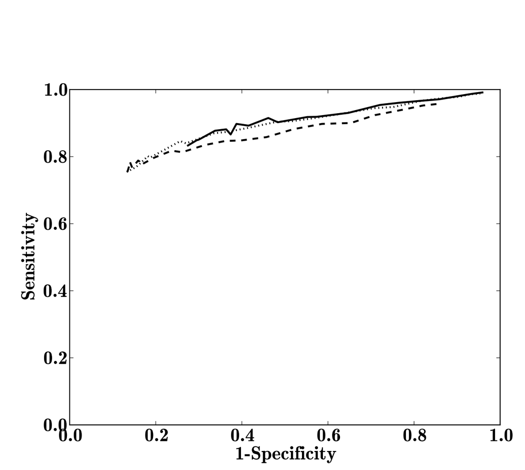
Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...



(a)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[1]

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[6] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[23] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[24] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

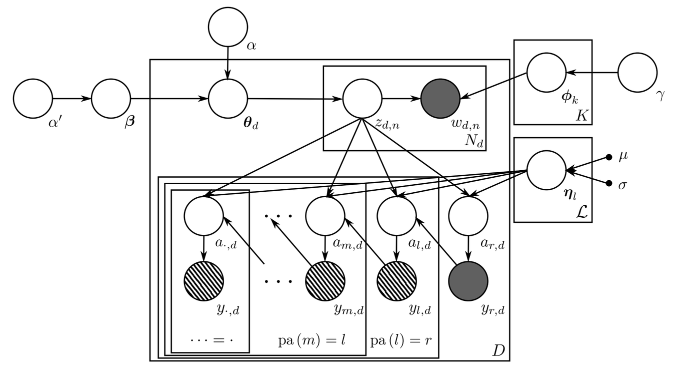


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [**? ?**]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [28]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 12, 19, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [13], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied after inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

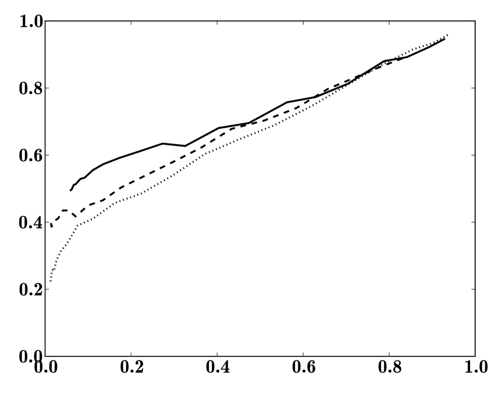
Figures 2 and 3 show the predictive performance of HSLDA relative to the three comparison models.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

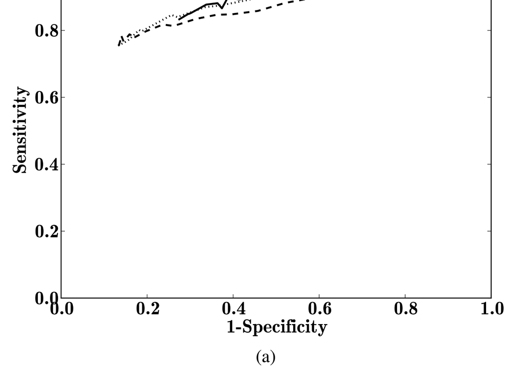
Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...



(a)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[1]

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[6] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[23] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[24] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://portal.acm.org/citation.cfm?id=1699510.1699543`.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

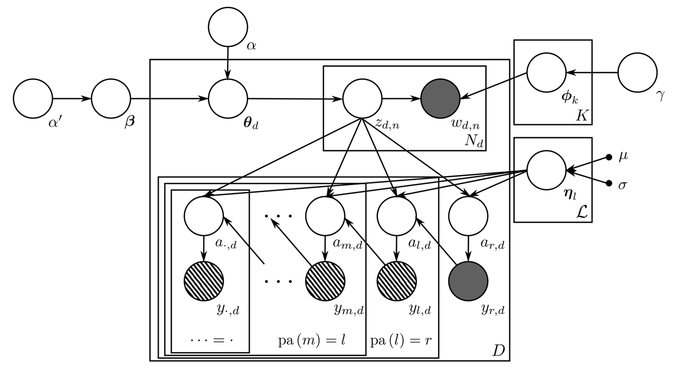


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [16]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, η the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. β is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling β is done using the "direct assignment" method of Teh et al. [28]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β. Their conditional posterior distribution is sampled according to

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 12, 19, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient's complaints, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [13], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics of performance - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that model performance generalizes over two very disparate domains.

To evaluate the performance of these models, we establish a gold standard for comparison. In our evaluation of each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by focusing on the observed labels as positives. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied after inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

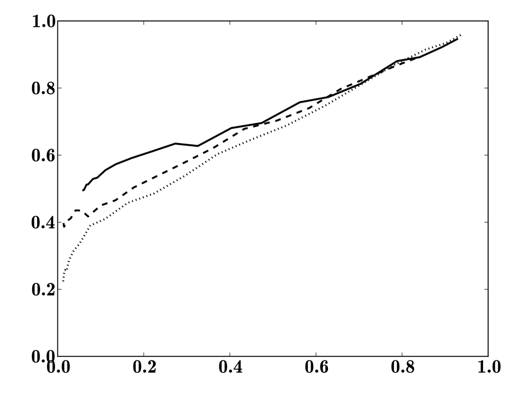
Figures 2 and 3 show the predictive performance of HSLDA relative to the three comparison models.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

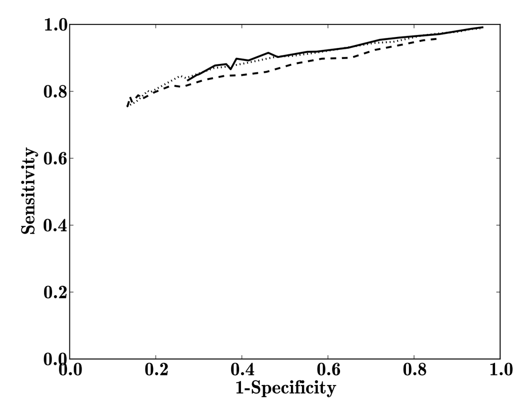
Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for exceptional purposes.

...



(a)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.



(a)

Figure 3: Out-of-sample Amazon product category predictions from product free-text descriptions. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** shows predictive performance as a function of the auxiliary variable threshold and (a) shows predictive perfmance as a function of the prior mean on regression parameters.

References

[1] Amazon, Inc. http://www.amazon.com/, 2011.

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[6] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] Yan Yan David S. Nilasena Martha J. Radford Brian F. Gage Elena Birman-Deych, Amy D. Waterman. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[14] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] Mehrdad Farzandipour, Abbas Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] Leah Larkey and Bruce Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[23] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[24] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://portal.acm.org/citation.cfm?id=1699510.1699543.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multi-labeled bag-of-word data. Examples of such data include Web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs (e.g. [1] as available from [5]), and patient hospital treatment transcripts and codes applied to them for bookkeeping and insurance purposes (e.g. hospital discharge records with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to automatically categorize new text documents, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

Our main contribution is to show how to utilize supervision in the form of hierarchical and (often) multiple labelings in a similar manner. Consider web retail data. Web retailers often have both a browse-able product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling of the free-text product descriptions. We hypothesize that such hierarchical labels should, at least in theory, provide better supervision than the simpler unstructured labels previously considered. Results from applying our model to both medical record and web retail data suggests that this is likely the case. In particular, we observe gains in our primary goal of out-of-sample label prediction that result specifically from leveraging hierarchical supervision.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In

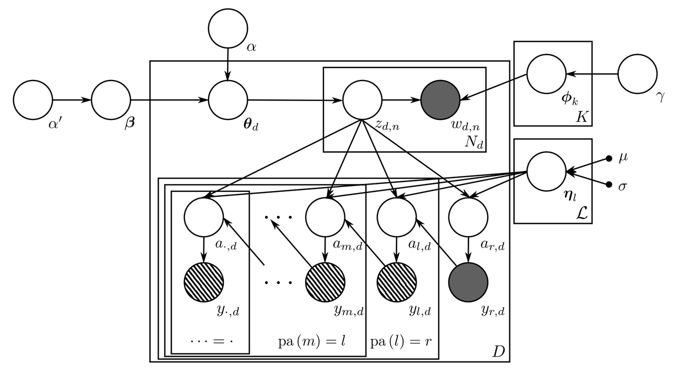


Figure 1: HSLDA graphical model

it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [16]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured textual data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no negative labeling present in the data naturally, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [8] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering unstructured labels.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to

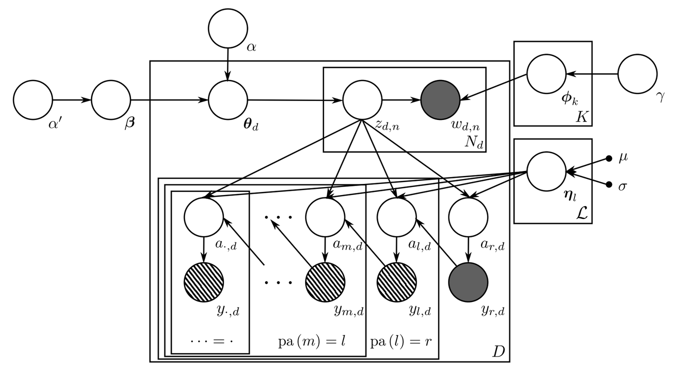


Figure 1: HSLDA graphical model

every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{1}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{1}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha \beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [16]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, η the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \tag{2}$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[6]

$$p(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. β is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling β is done using the "direct assignment" method of Teh et al. [28]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma(\alpha \beta_k)}{\Gamma\left(\alpha \beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) (\alpha \beta_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [8] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [9] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [8]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [8]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 11] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 13, 19, 10].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [7], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [12, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

been shown to improve performance in sLDA [8]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

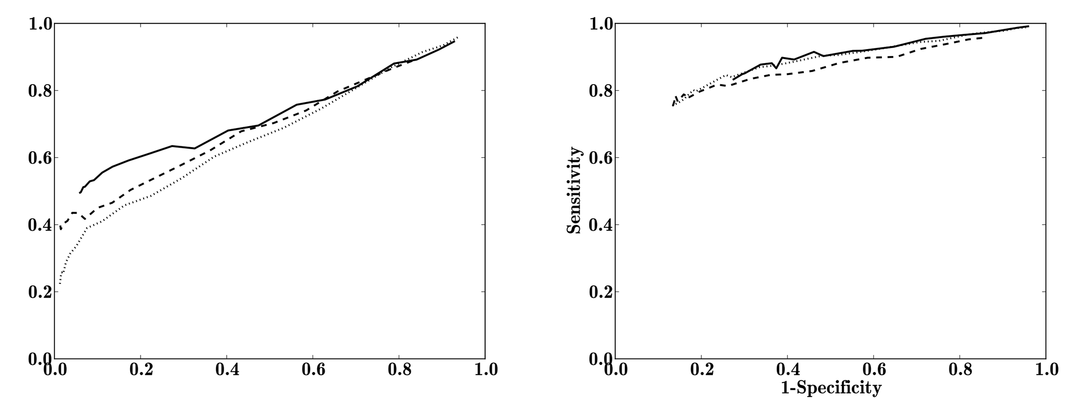
To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure **??** shows the predictive performance of HSLDA relative to the three comparison models. In figure **??**.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (**??** and **??**). Out-of-sample Amazon product category predictions from product free-text descriptions (**??** and **??**). In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** and **??** show predictive performance as a function of the auxiliary variable threshold and **??** and **??** show predictive perfmance as a function of the prior mean on regression coefficients.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] Amazon, Inc. `http://www.amazon.com/`, 2011.

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[6] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[8] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[10] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[11] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[12] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[13] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[14] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[23] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[24] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [3], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [4]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on ISA hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [8] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \dots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

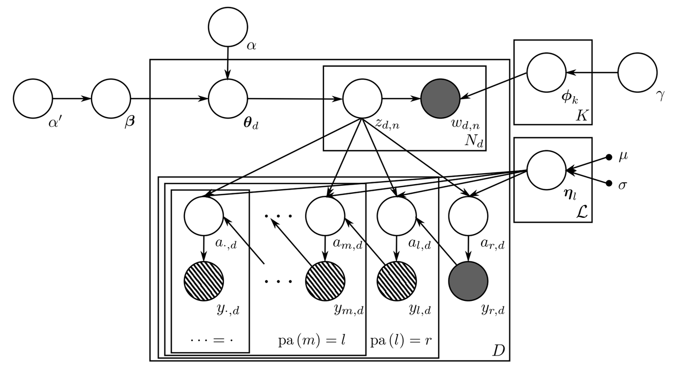


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \dots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol\phi_k$ is a distribution over "words," $\boldsymbol\theta_d$ is a document-specific distribution over topics, $\boldsymbol\beta$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \dots, K$
   - Draw a distribution over words $\boldsymbol\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \dots, D$
   - Draw topic proportions $\boldsymbol\theta_d \mid \boldsymbol\beta, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol\beta)$
   - For $n = 1, \dots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol\theta_d \sim \mathrm{Multinomial}(\boldsymbol\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol\phi_{1:K} \sim \mathrm{Multinomial}(\boldsymbol\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol\eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol\eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol\eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \dots, \bar{z}_k, \dots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [16]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol\eta_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [18], we have analytically marginalized out the parameters $\boldsymbol\phi_{1:K}$ and $\boldsymbol\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol\eta$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol\eta, \alpha, \boldsymbol\beta, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol\eta_l - a_{l,d})^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \dots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [6]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [6].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol\eta) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol\eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol\beta$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [29]. Sampling $\boldsymbol\beta$ is done using the "direct assignment" method of Teh et al. [28]

$$\boldsymbol\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \dots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol\beta$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol\beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [8] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [9] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generated as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [8]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [8]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[25] and DiscLDA[20]. Various applications of these models to computer vision and document networks have been explored [30, 11] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[23, 13, 19, 10].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [2], tend to be more specific than sensitive in their assignments [7], and sometimes make mistakes [15].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [2]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [12, 17, 14]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [21, 26, 24, 27, 22].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based search or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [5]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additional structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

been shown to improve performance in sLDA [8]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \dots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol\eta_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

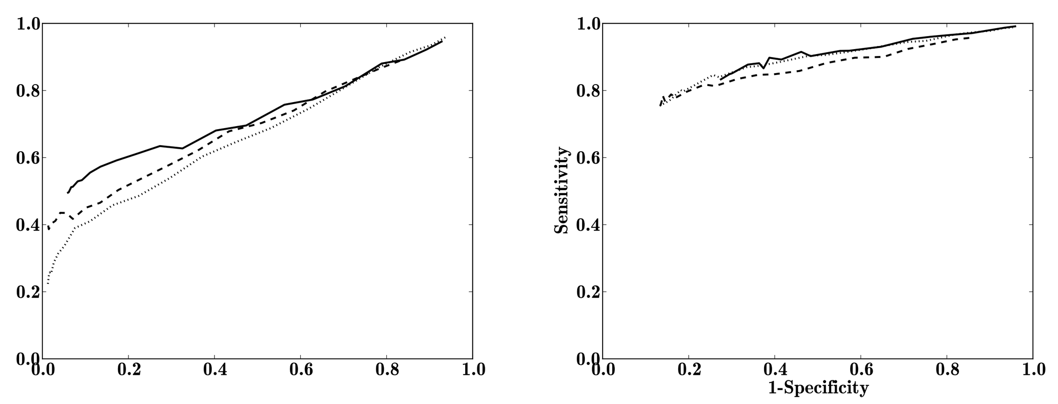
To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models. In figure 2(a).

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (**??** and (a)). Out-of-sample Amazon product category predictions from product free-text descriptions (**??** and (b)). In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** and **??** show predictive performance as a function of the auxiliary variable threshold and (a) and (b) show predictive perfmance as a function of the prior mean on regression coefficients.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[2] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[3] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[5] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[6] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[7] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[8] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[10] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[11] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[12] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[13] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[14] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[16] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[17] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[22] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[23] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[24] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[25] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[26] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[27] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[29] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no negative labeling present in the data naturally, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. In particular, we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We hypothesize that hierarchical label information provides more information about labeling than considering unstructured labels.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to

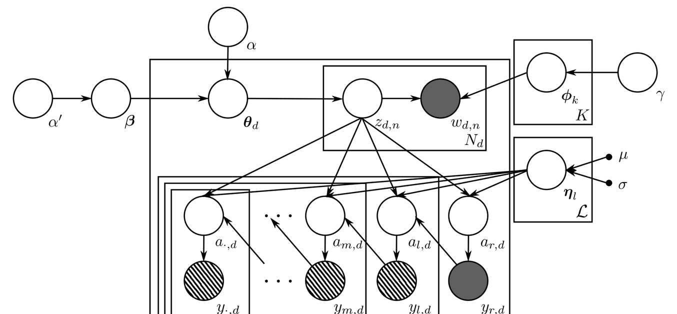


Figure 1: HSLDA graphical model

every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[5]

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the predictive performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models. In figure 2(a).

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

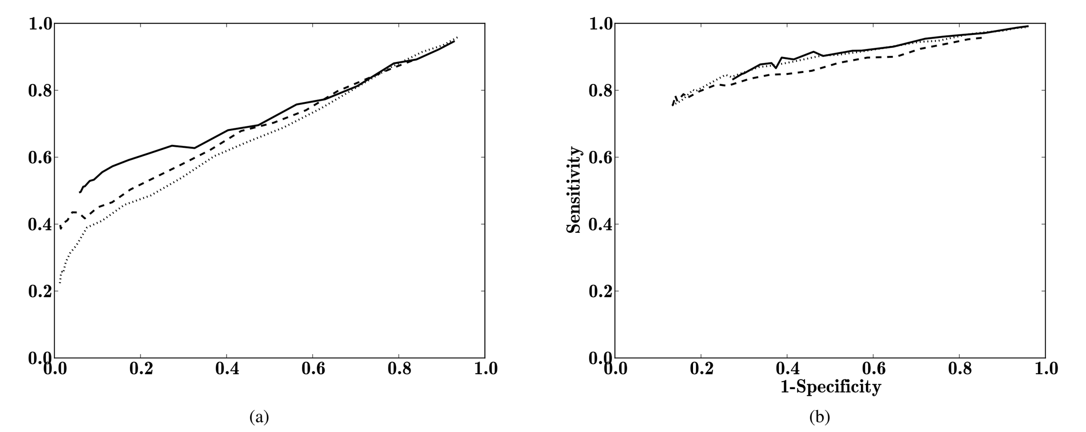


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (?? and (a)). Out-of-sample Amazon product category predictions from product free-text descriptions (?? and (b)). In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. ?? and ?? show predictive performance as a function of the auxiliary variable threshold and (a) and (b) show predictive perfmance as a function of the prior mean on regression coefficients.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures ?? and 2(a) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on ISA hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. The set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

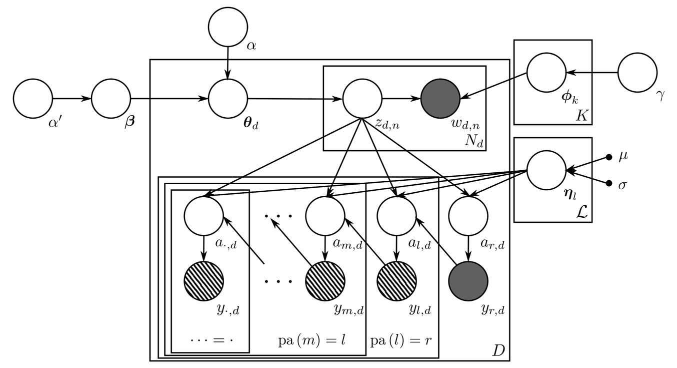


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol\phi_k$ is a distribution over "words," $\boldsymbol\theta_d$ is a document-specific distribution over topics, $\boldsymbol\beta$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol\theta_d \mid \boldsymbol\beta, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol\beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol\theta_d \sim \mathrm{Multinomial}(\boldsymbol\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol\phi_{1:K} \sim \mathrm{Multinomial}(\boldsymbol\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol\eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1}\sum_{n=1}^{N_d}\mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol\eta_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol\phi_{1:K}$ and $\boldsymbol\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol\eta$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol\eta, \alpha, \boldsymbol\beta, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol\mu}_l, \hat{\boldsymbol\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol\mu}_l = \hat{\boldsymbol\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right) \qquad \hat{\boldsymbol\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[5]

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol\eta\right) \propto \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol\eta_l^T\bar{\mathbf{z}}_d\right)\right\}\mathbb{I}\left(a_{l,d}y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol\beta$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol\beta$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol\beta$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol\beta\right) = \frac{\Gamma\left(\alpha\boldsymbol\beta_k\right)}{\Gamma\left(\alpha\boldsymbol\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right)\left(\alpha\boldsymbol\beta_k\right)^m \quad (5)$$

where $s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision", often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generated as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based search or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol\eta_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models. In figure 2(a).

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

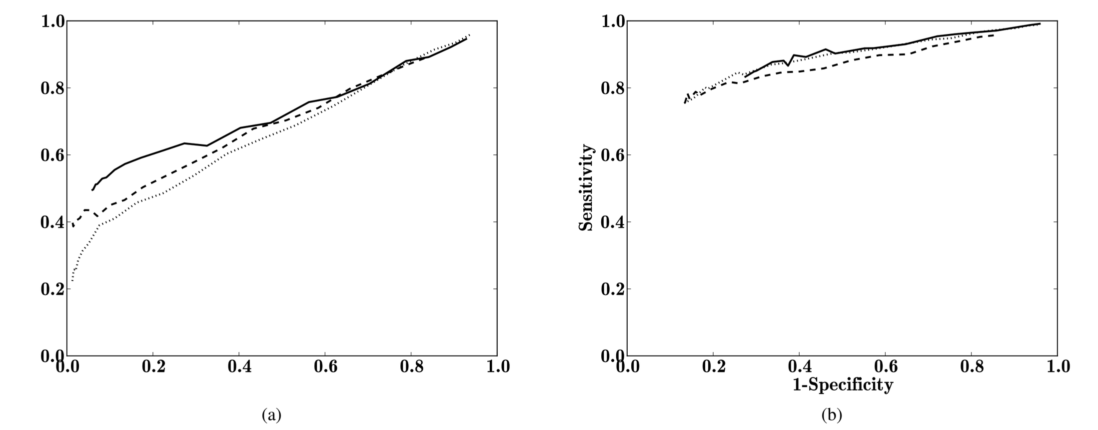


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (**??** and (a)). Out-of-sample Amazon product category predictions from product free-text descriptions (**??** and (b)). In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** and **??** show predictive performance as a function of the auxiliary variable threshold and (a) and (b) show predictive perfmance as a function of the prior mean on regression coefficients.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures **??** and 2(a) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on ISA hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

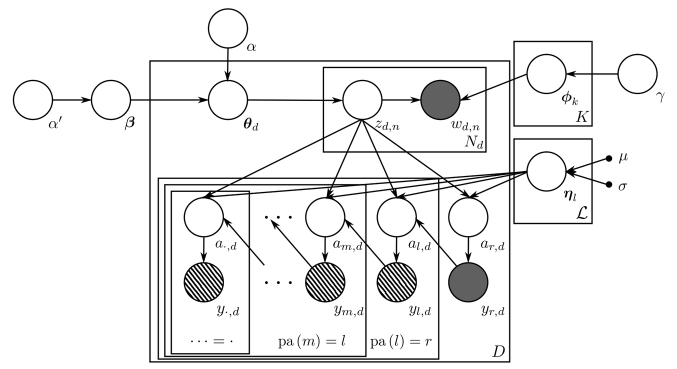


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[5]

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) \left(\alpha\boldsymbol{\beta}_k\right)^m \tag{5}$$

where $s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complaints, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based search or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure ?? shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficient. Figure ?? demo.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

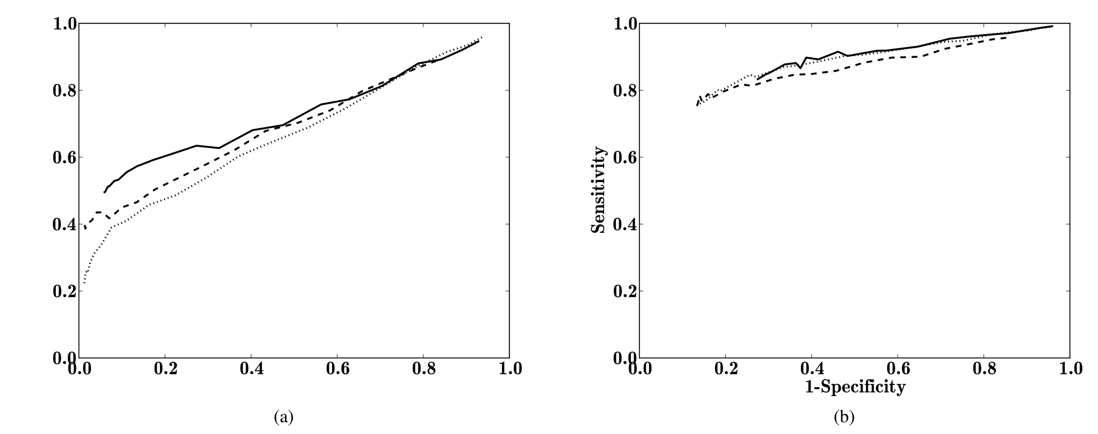


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (??). Out-of-sample Amazon product category predictions from product free-text descriptions (??). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures ?? and ?? suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition.* Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g. image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other structures. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except the root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

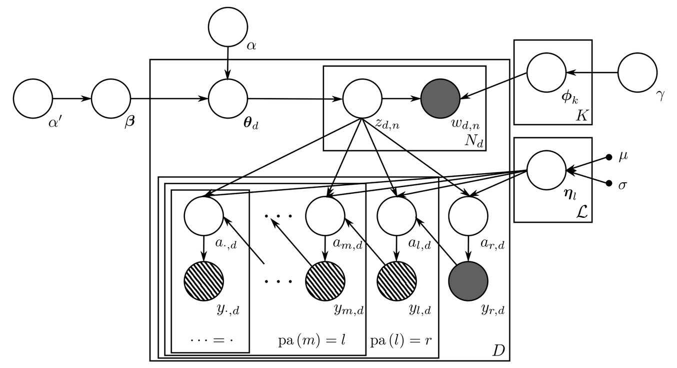


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \bar{\mathbf{z}}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right)(\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision", often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of New-York-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based search or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

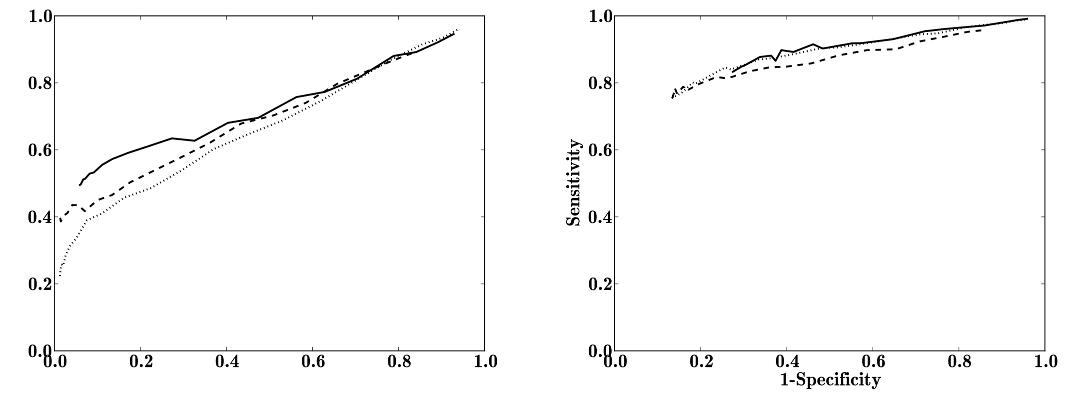
To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models. In figure 2(b).

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records (**??** and (b)). Out-of-sample Amazon product category predictions from product free-text descriptions (**??** and **??**). In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions. **??** and **??** show predictive performance as a function of the auxiliary variable threshold and (b) and **??** show predictive perfmance as a function of the prior mean on regression coefficients.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data is also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured textual data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other domains as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on ISA hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-words data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label except root labels $l \in \mathcal{L}$ has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more complicated inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

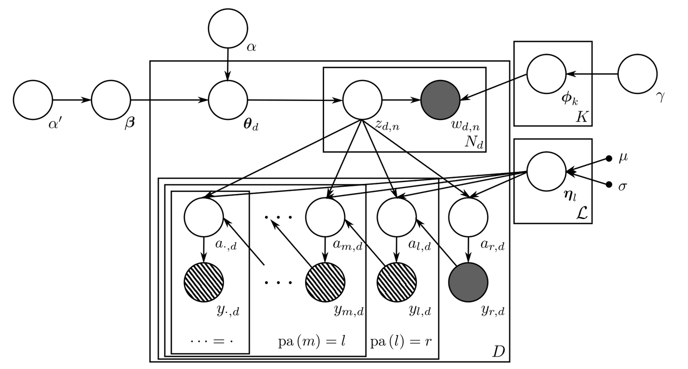


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5]

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision", often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), ratings given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complaints, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of New York-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficient. Figure **??** demo.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

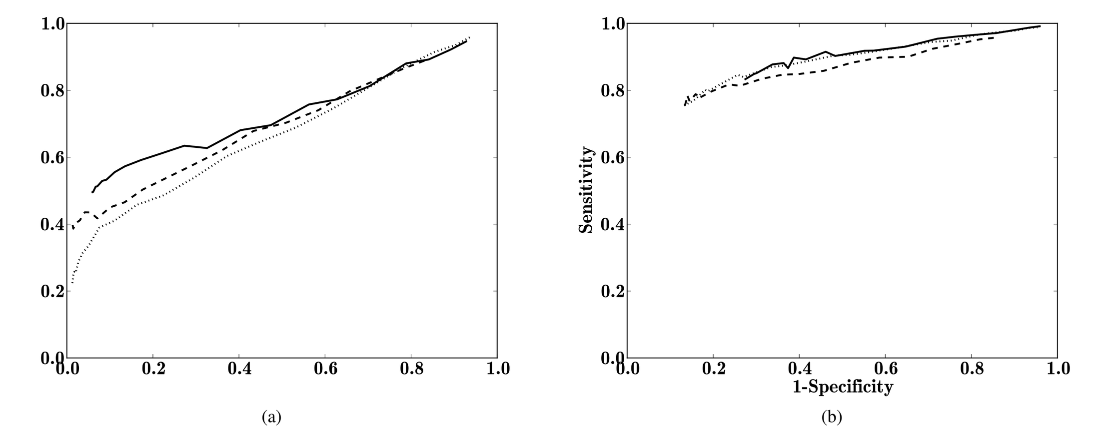


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((a)). Out-of-sample Amazon product category predictions from product free-text descriptions ((a)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

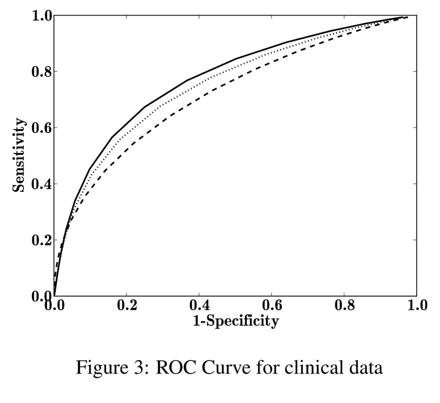


Figure 3: ROC Curve for clinical data

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures **??** and 2(a) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except the root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{l,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

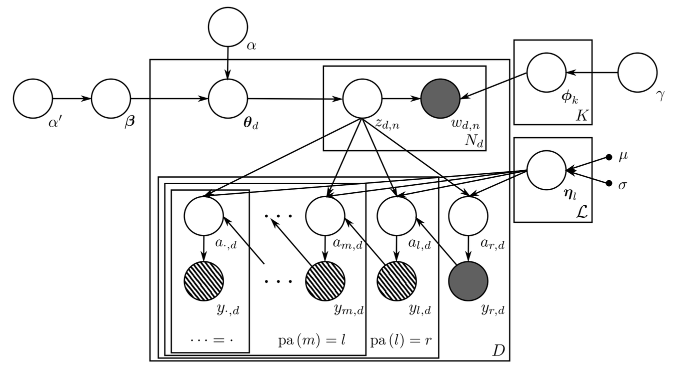


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e. $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e. $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, the bag-of-words document data is modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In it and the following K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here each document is generatively labeled using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the readers attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto$$

$$\left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution[5]

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e. $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catolog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficient. Figure 2(a) demo.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

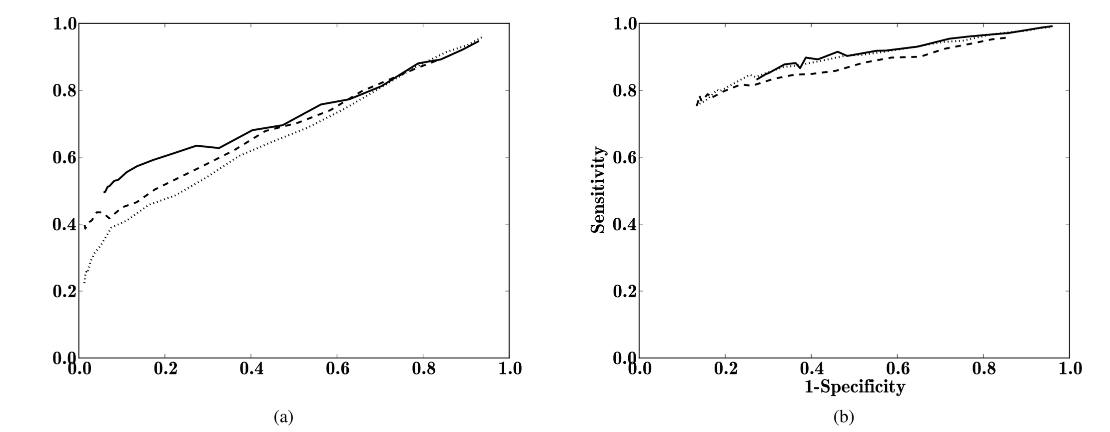


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

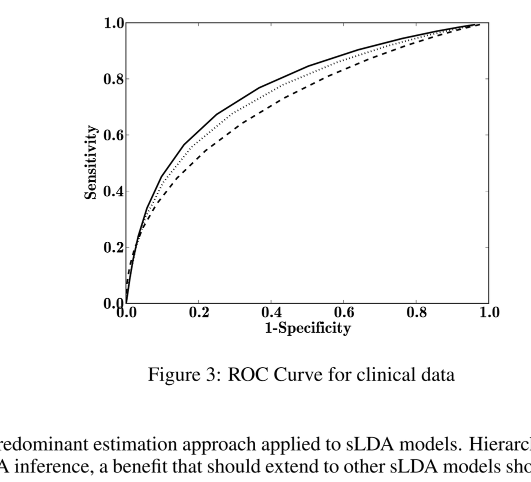


Figure 3: ROC Curve for clinical data

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

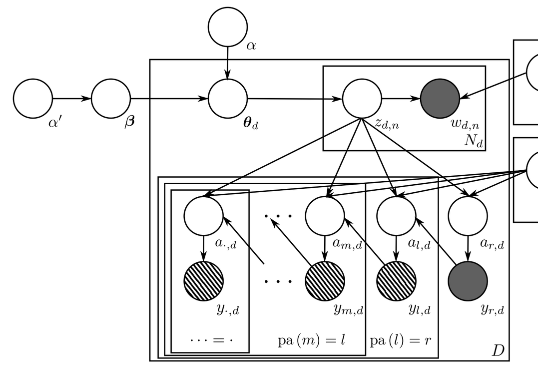


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], and tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additional structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

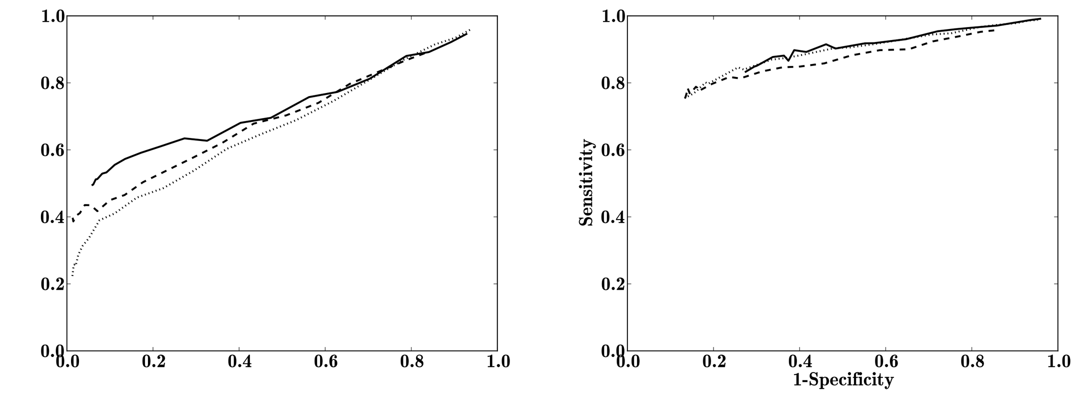
To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficient. Figure 2(a) demo.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

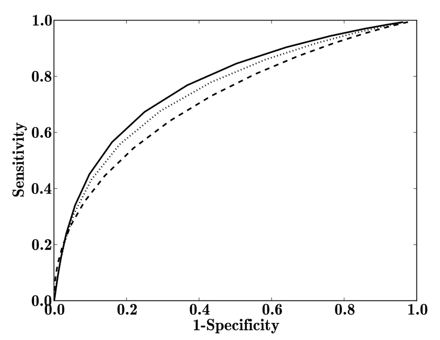


Figure 3: ROC Curve for clinical data

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

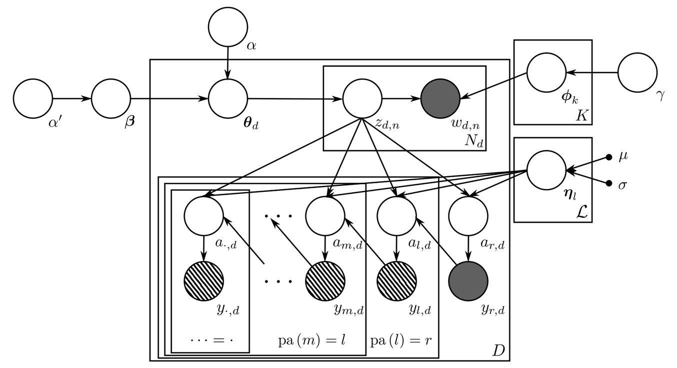


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1}\sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right)(\alpha\beta_k)^m \quad (5)$$

where $s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision"; often taking the form of a single numerical or categorical label. More generally this supervision is just extra per-document data; for instance its quality or relevance (e.g. online review scores), marks given to written work (e.g. essay grades), or the number of times a web page is linked. These labels are usually generatively modeled as a conditional draw from some distribution that depends on each document-specific topic mixture. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we contrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA[22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can,

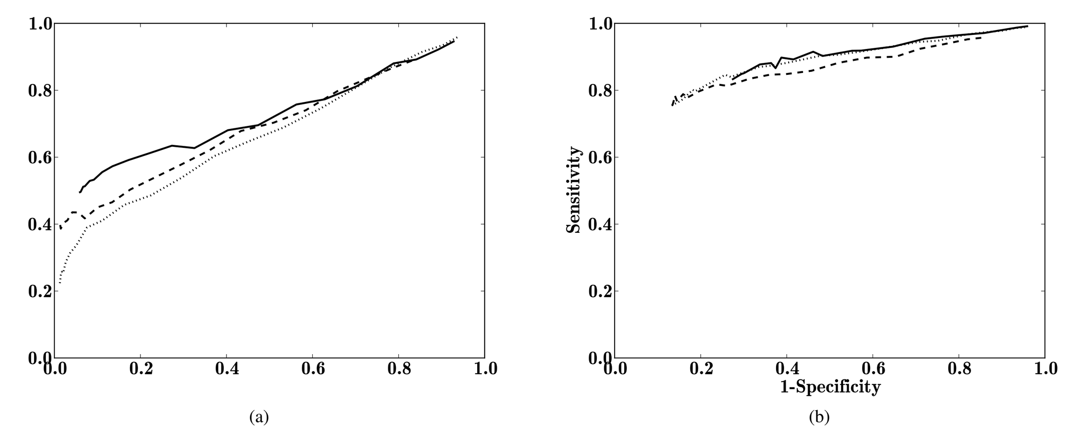


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

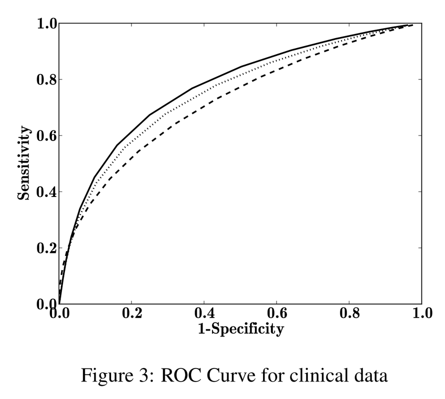


Figure 3: ROC Curve for clinical data

in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \dots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

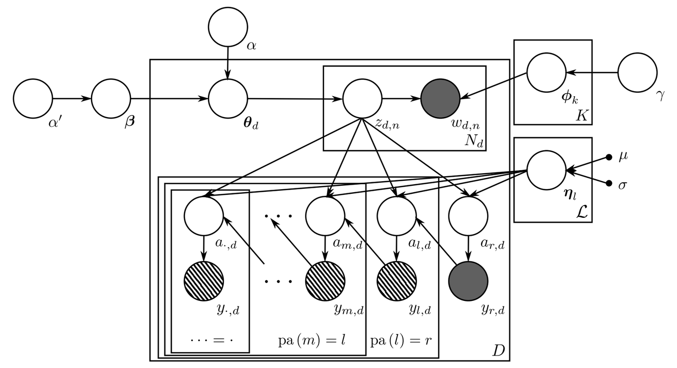


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \dots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{1}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \dots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \dots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \dots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \dots, \bar{z}_k, \dots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \dots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \dots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We experimented with HSLDA for prediction in two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize the course of a hospitalized patient. The summaries typically contain a record of the patient complains, findings and diagnoses, along with treatment and hospital course. For each admission trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] As such, the ICD-9 codes constitute a labeling of a patient's diagnoses based on a discharge summary. The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Much of the work was triggered by the 2007 medical NLP community challenge [1]. The data in the challenge, however, differs from ours in its scope. The datasets were smaller (1,000 training and 1,000 testing documents) and focused on radiology reports with a restricted number of ICD-9 codes (45 of them, compared to 7K+ in our dataset). Methods ranged from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of NewYork-Presbyterian Hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] Vocabulary was determined as the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). Each discharge summary is thus represented as counts over the 10,000-word vocabulary. The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a hierarchy and provides product descriptions for most products in their catalog. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

We were able to deduce the structure of the hierarchy for the Amazon.com products because all ancestors in the hierarchy were included with each category label. For example, "DVD / Genres / Science Fiction & Fantasy / Classic Sci-Fi" is a single product category for the DVD, "The Time Machine."

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by performing LDA followed by tree-conditional regressions, and HSLDA fit with fixed random regression parameters. These models were chosen to highlight several aspects of the model including performance in the absence of hierarchical constraints, the effect of the combined inference procedure, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additional structure imposed on the label space, a distinction that we hypothesize will result in a difference in predictive performance. Aside from this, all other features of sLDA are preserved between the two models.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

The last comparison model is HSLDA with fixed and randomly selected regression parameters. There is a baseline benefit that the structure in the label space provides for the prediction of labels. This comparison model is intended to quantify the contribution of the structure alone.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \dots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. Specifically, ancestors of observed nodes were ignored, observed nodes were considered positive and unobserved nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficient. Figure 2(a) demo.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

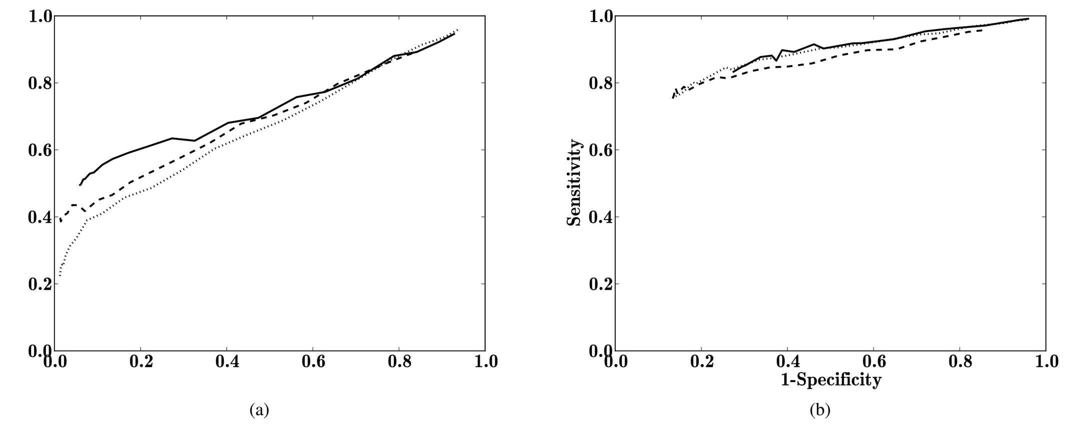


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

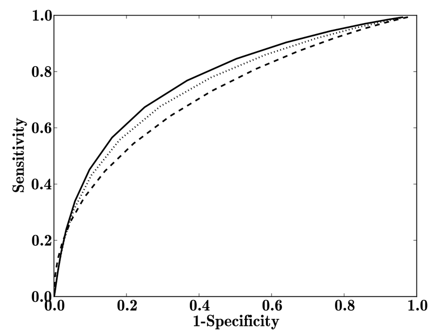


Figure 3: ROC Curve for clinical data

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

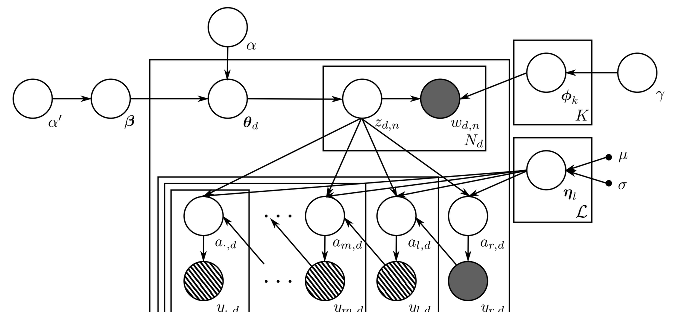


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \tag{1}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha\beta_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a multiply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

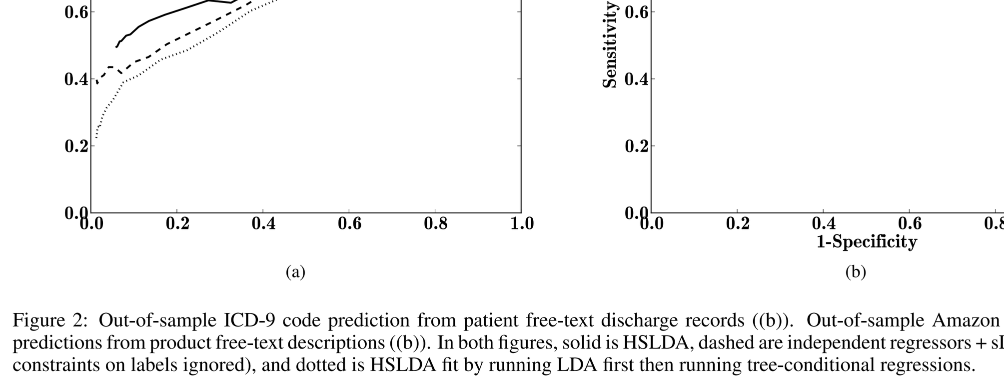
Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can,



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

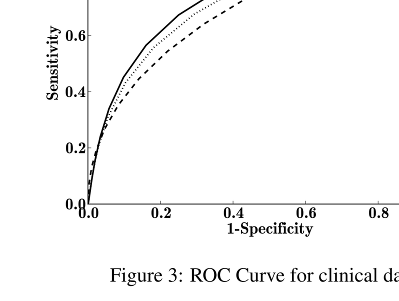


Figure 3: ROC Curve for clinical data

in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

...

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

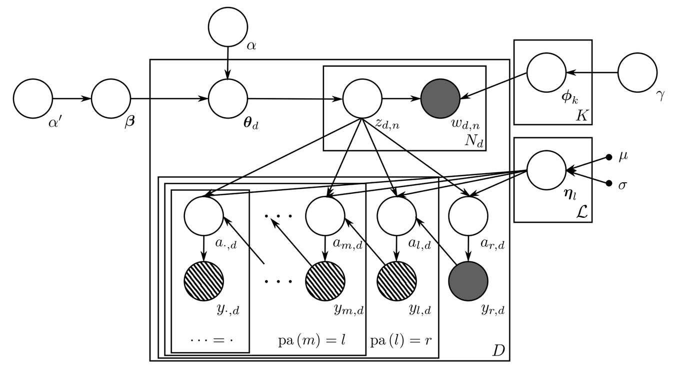


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," θ_d is a document-specific distribution over topics, β is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\theta_d \mid \beta, \alpha \sim \mathrm{Dir}_K(\alpha\beta)$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \theta_d \sim \mathrm{Multinomial}(\theta_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \phi_{1:K} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients η_l are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\phi_{1:K}$ and $\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, η the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \eta, \alpha, \beta, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\beta_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \eta_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \eta\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \eta_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., β is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling β is done using the "direct assignment" method of Teh et al. [27]

$$\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of β. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \beta\right) = \frac{\Gamma\left(\alpha\beta_k\right)}{\Gamma\left(\alpha\beta_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) \left(\alpha\beta_k\right)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a multiply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[6]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and η_l were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

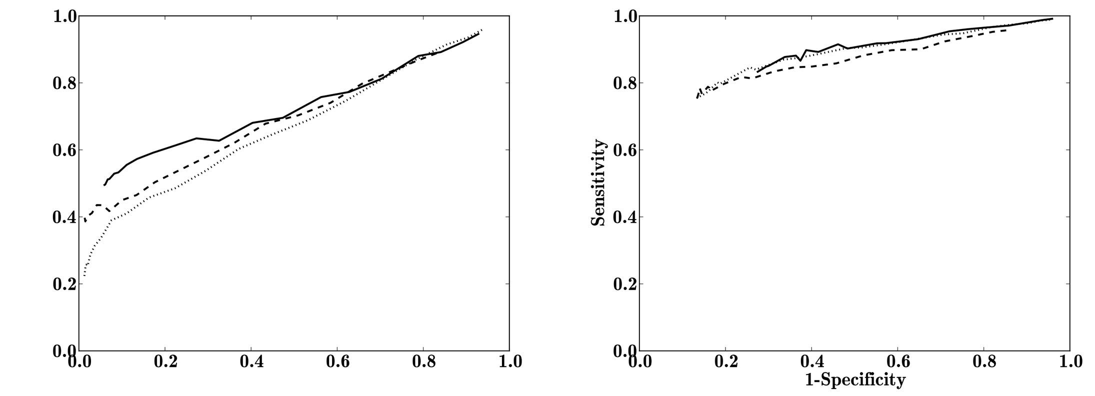
For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. Another way to see the family is as a set of models that can predict labels for bag-of-words data. A large diversity of problems can be expressed as label prediction problems for bag-of-words data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

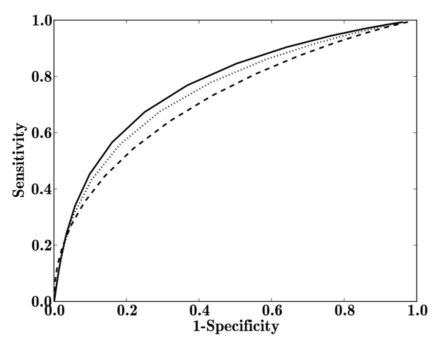


Figure 3: ROC Curve for clinical data

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. In alternative interpretation of the same results is that if one is more sensitive to the performance gains that result from exploiting the structure of the labels then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other structures as well. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

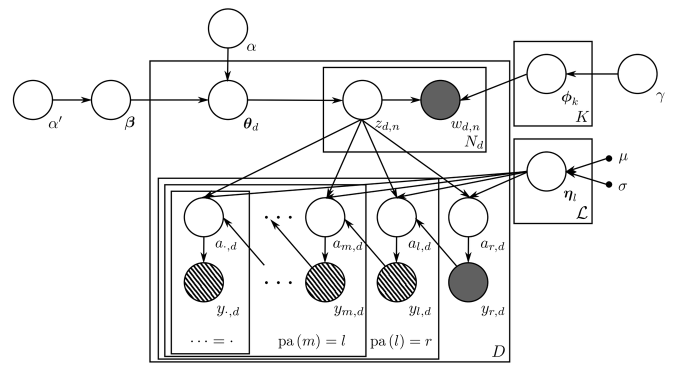


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \quad (5)$$

where $s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through query-based searching or through product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance. Predictive performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate). We evaluate on the two aforementioned datasets to demonstrate that our model generalizes to two different domains.

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, the models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely lead to a slight overestimation of the number of false positives. It is known that ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ where $\hat{d}$ represents the test document. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set.

The following are the predictive performance results for the clinical data given a prior mean for the regression parameters of -1.6. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of -2.2. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

To further explore this tradeoff between the true positive rate and the false positive rate we evaluated predictive performance for a range of values for two different parameters - the prior mean for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one

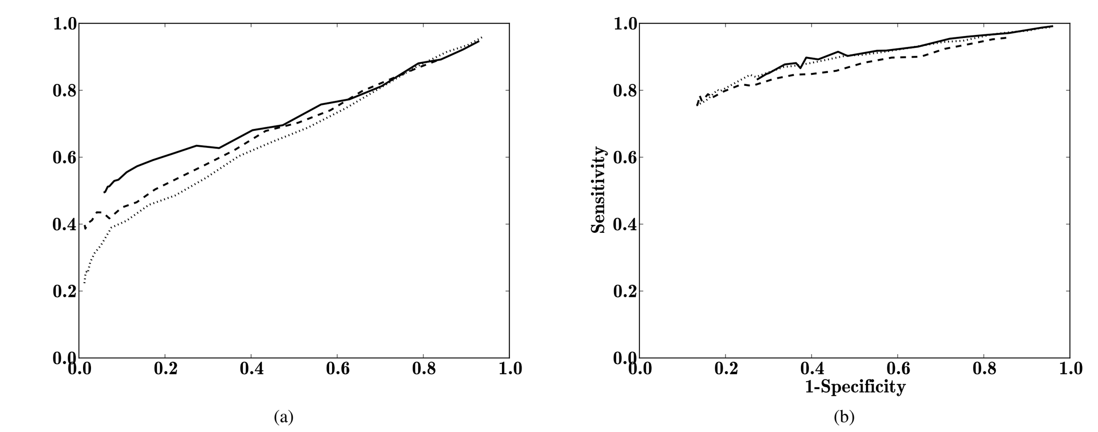


Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

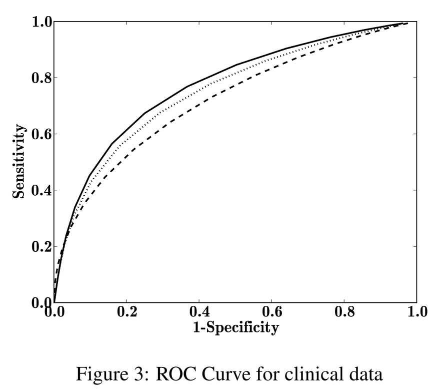


Figure 3: ROC Curve for clinical data

can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable in other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

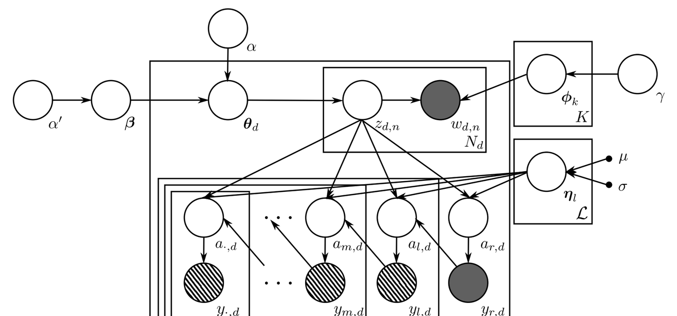


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a multiply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[6]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

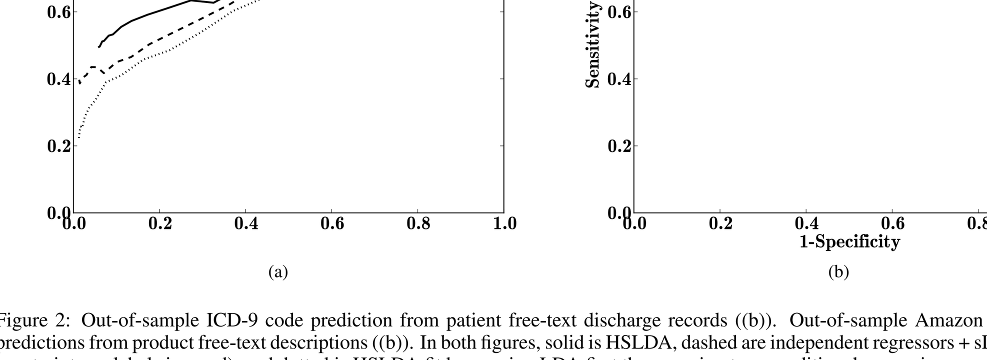
For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

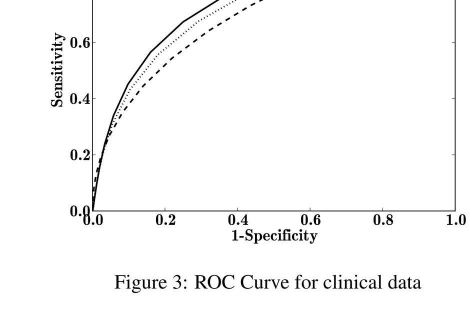


Figure 3: ROC Curve for clinical data

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

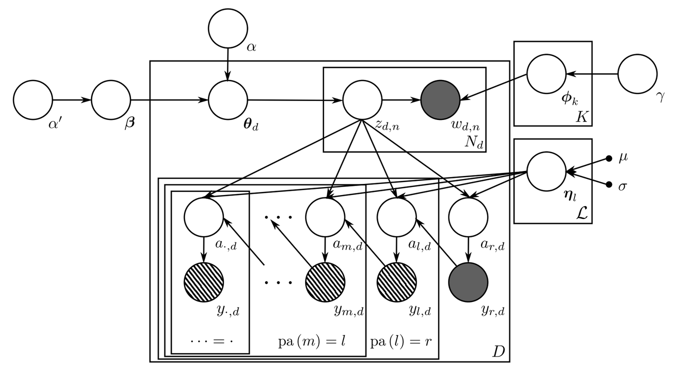


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 26, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[6]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

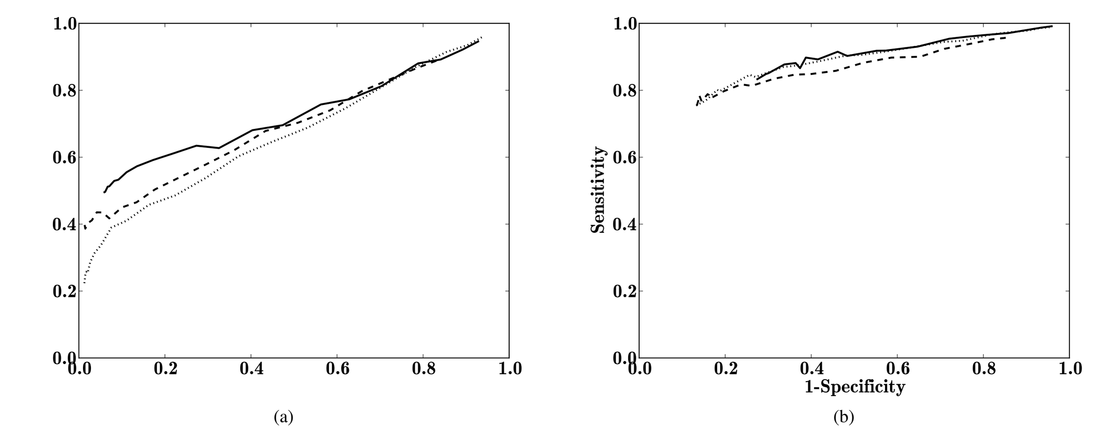
For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

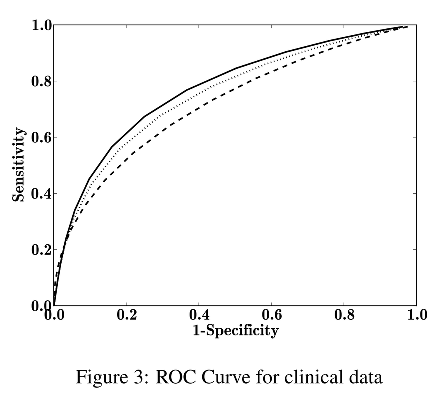


Figure 3: ROC Curve for clinical data

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachusetts, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes (e.g. hospital discharge summaries with International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes assigned [3]). In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except the root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

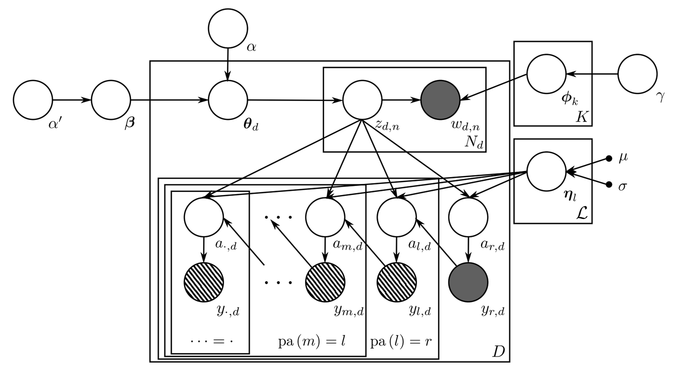


Figure 1: HSLDA graphical model

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \bar{\mathbf{z}}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [28]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [27]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [29, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[6]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

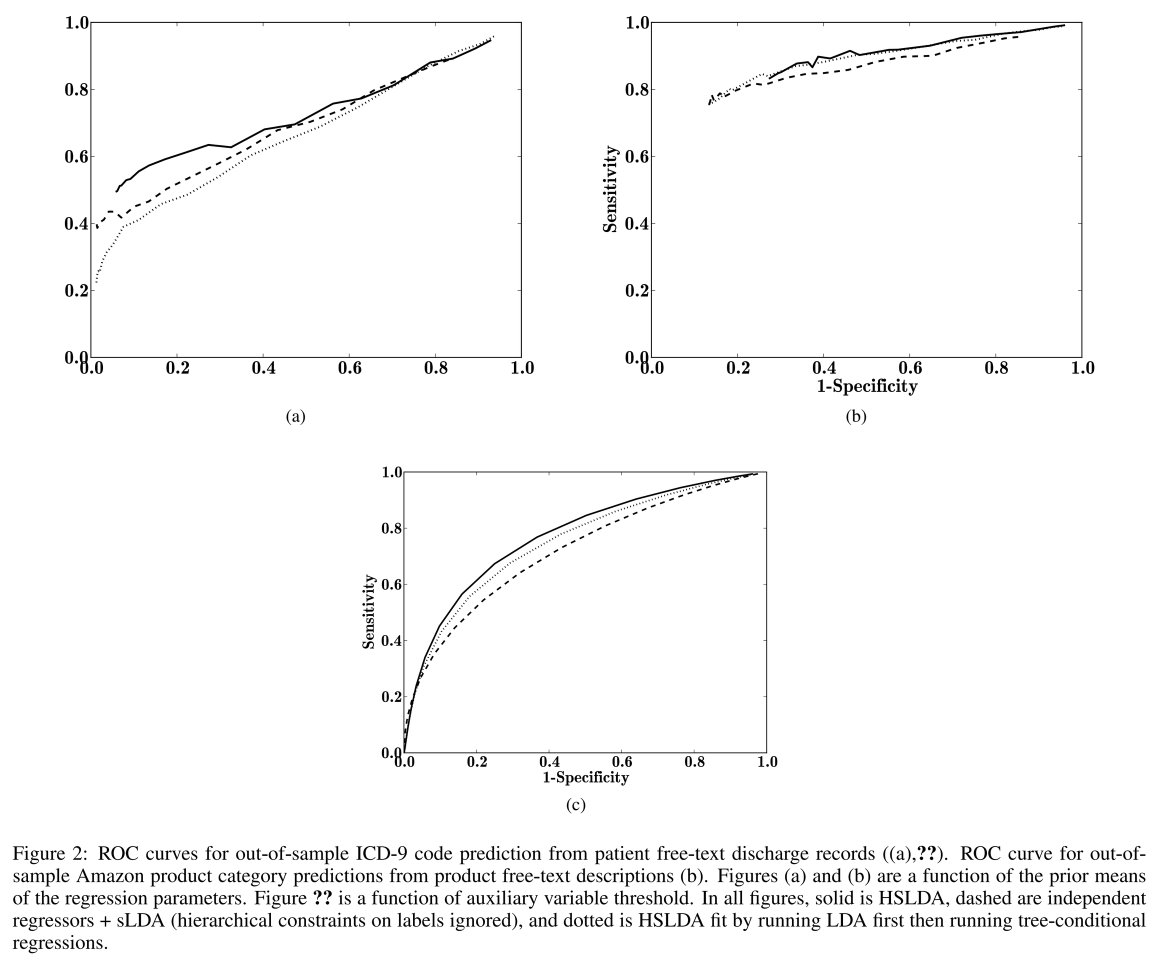


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),**??**). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure **??** is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[4] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[28] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

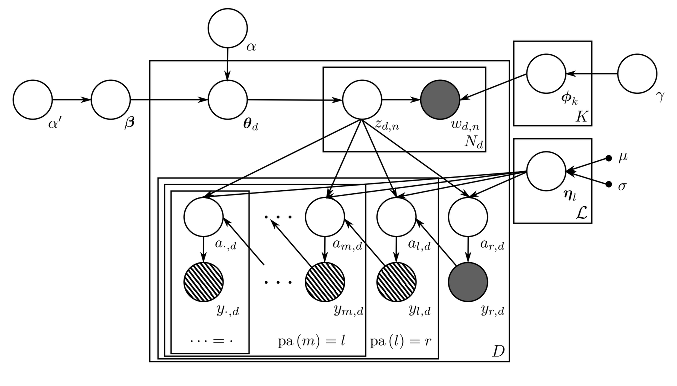


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [28, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three other closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm

[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

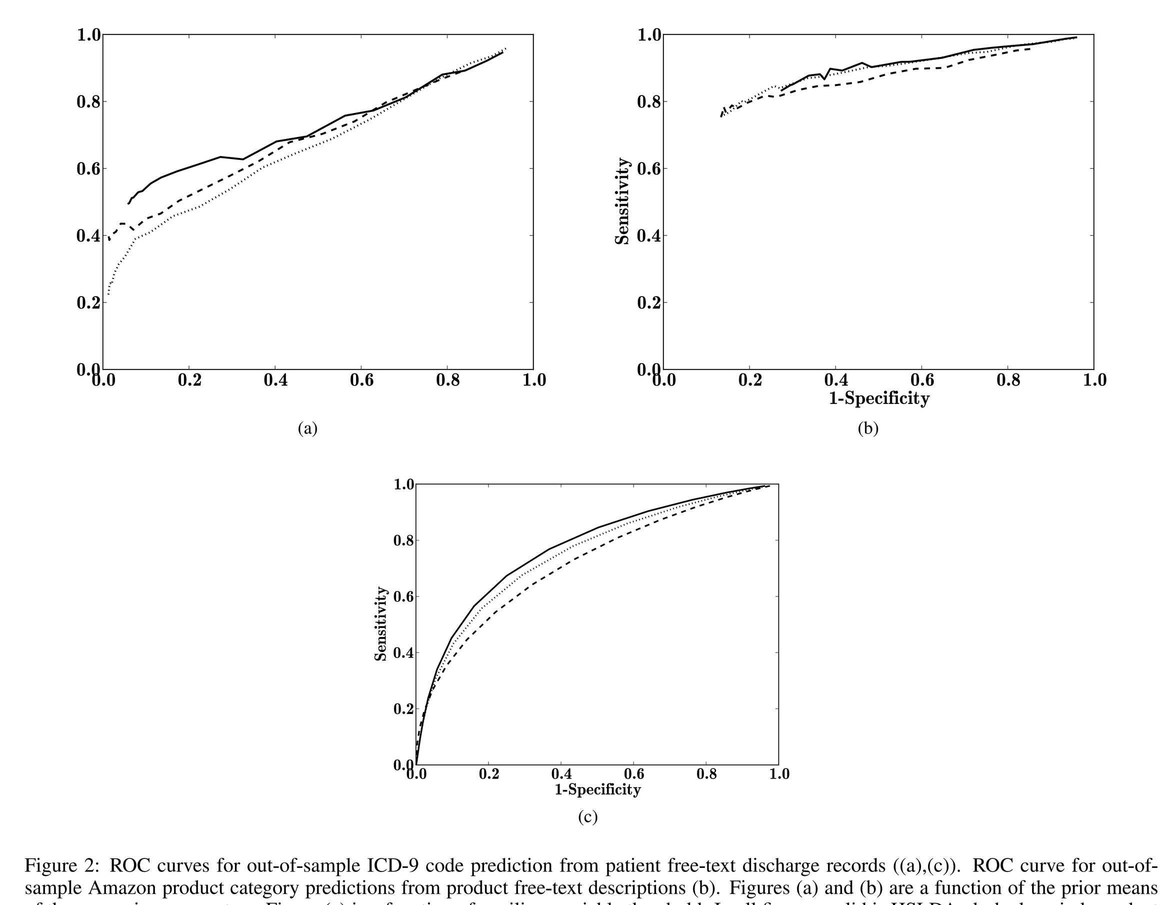


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[] *ICD-9-CM: international classification of diseases, 9th revision; clinical modification, 6th edition*. Practice Management Information Corporation, Los Angeles, CA, 2006.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

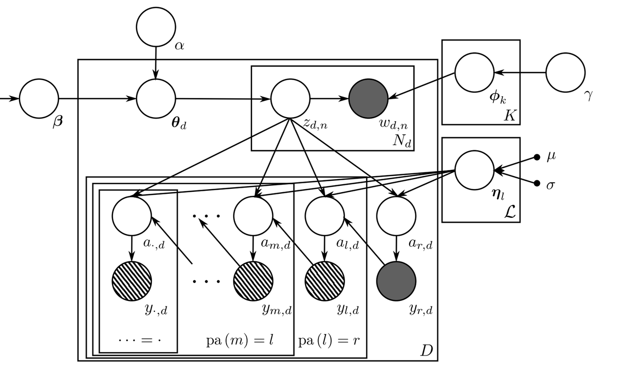


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [15]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [17], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha \boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [5]; the specific form of the update is a standard result from Bayesian normal linear regression [15]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

$$p\left(a_{l,d}, \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [7] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [8] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [7]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [7]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[24] and DiscLDA[19]. Various applications of these models to computer vision and document networks have been explored [28, 10] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [22, 12, 18, 9].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [14].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [11, 16, 13]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [20, 25, 23?, 21].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [4]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [7]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitivity (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[6]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

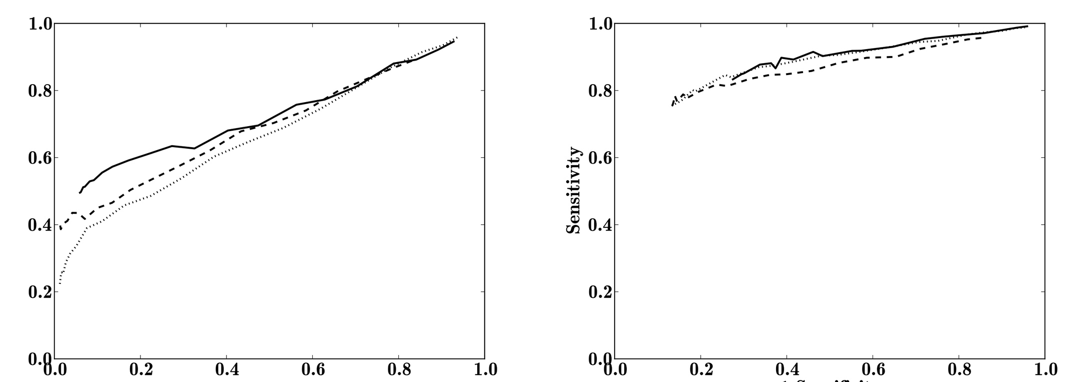
For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.



(a) (b)

Figure 2: Out-of-sample ICD-9 code prediction from patient free-text discharge records ((b)). Out-of-sample Amazon product category predictions from product free-text descriptions ((b)). In both figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

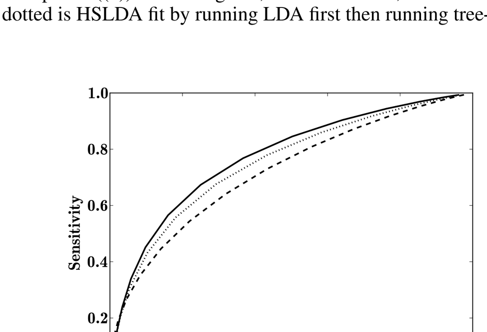


Figure 3: ROC Curve for clinical data

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[4] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[5] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[6] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[9] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[11] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[12] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[13] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[14] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[15] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[16] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[19] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[20] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[21] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[22] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[23] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[25] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

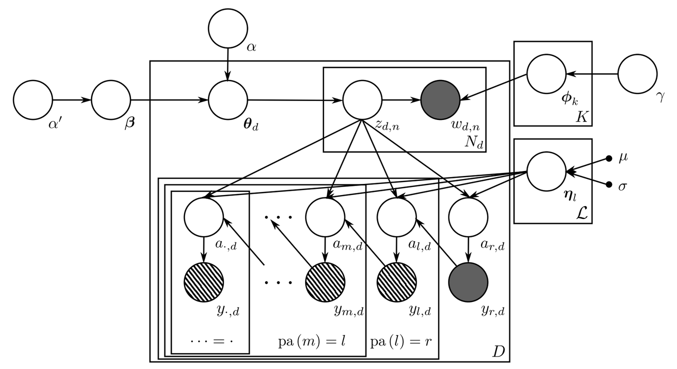


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), ϕ_k is a distribution over "words," $\boldsymbol\theta_d$ is a document-specific distribution over topics, $\boldsymbol\beta$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol\phi_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol\eta_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol\beta \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol\theta_d \mid \boldsymbol\beta, \alpha \sim \mathrm{Dir}_K(\alpha\boldsymbol\beta)$
   - For $n = 1, \ldots, N_d$
     – Draw topic assignment $z_{n,d} \mid \boldsymbol\theta_d \sim \mathrm{Multinomial}(\boldsymbol\theta_d)$
     – Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol\phi_{1:K} \sim \mathrm{Multinomial}(\boldsymbol\phi_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     – Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol\eta_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     – Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1}\sum_{n=1}^{N_d}\mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol\eta_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol\phi_{1:K}$ and $\boldsymbol\theta_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol\eta$ the set of all regression coefficients, and $\mathbf{z}_d\backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d\backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol\eta, \alpha, \boldsymbol\beta, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right)\frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)}\prod_{l\in\mathcal{L}_d}\exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T\boldsymbol\eta_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol\eta_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol\mu}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol\mu}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T\mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T\bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol\eta\right) \propto \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol\eta_l^T\bar{\mathbf{z}}_d\right)\right\}\mathbb{I}\left(a_{l,d}y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol\beta$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol\beta$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol\beta \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol\beta$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol\beta\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)}s\left(c^k_{(\cdot),d}, m\right)(\alpha\beta_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [27, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with three closely related models against these two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all of these models, particular attention was payed to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. We evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics - sensitiviy (true positive rate) and 1-specificity (false positive rate).

To evaluate the performance of these models, we established a gold standard for comparison. For each dataset, a held out set of 1000 documents and labels were reserved for evaluation and predictive performance was evaluated against a standard derived from the observed labeling. To make the comparison as fair as possible, ancestors of observed nodes were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. This method of defining positive and negative labels was chosen to be as fair as possible to all models being compared. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol\eta_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

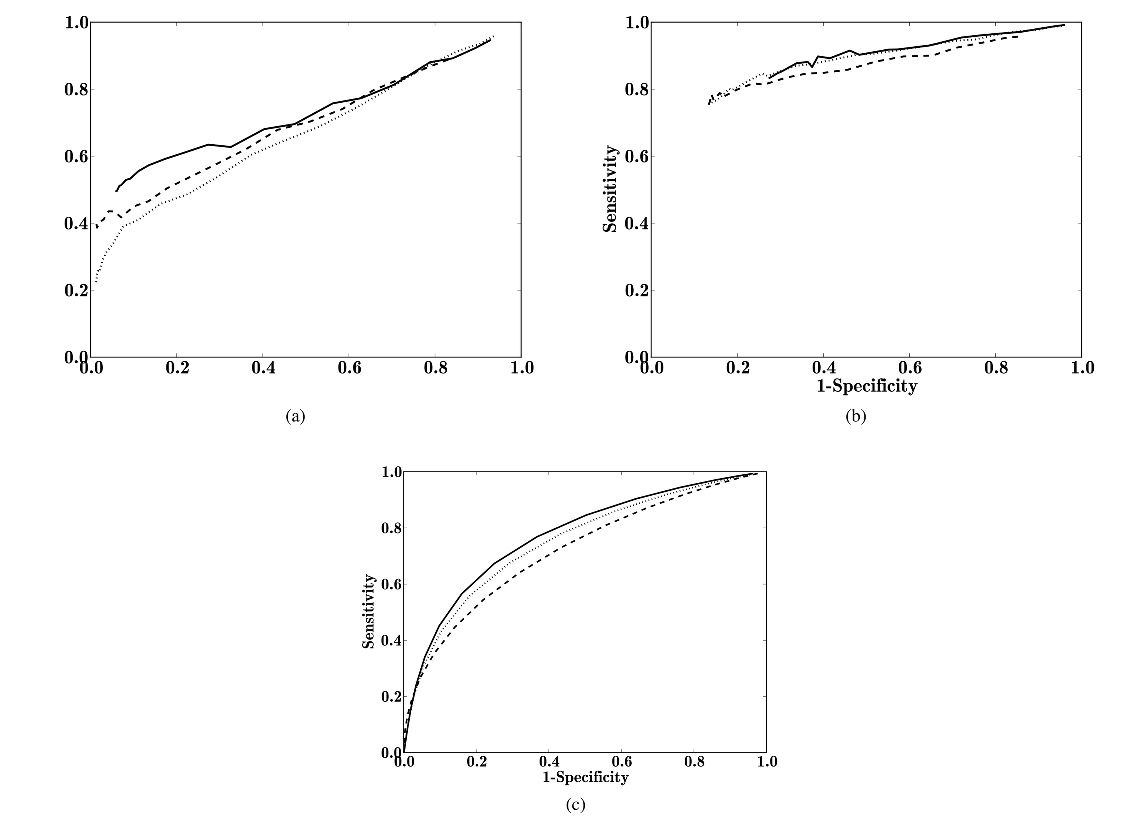


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),??). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure ?? is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

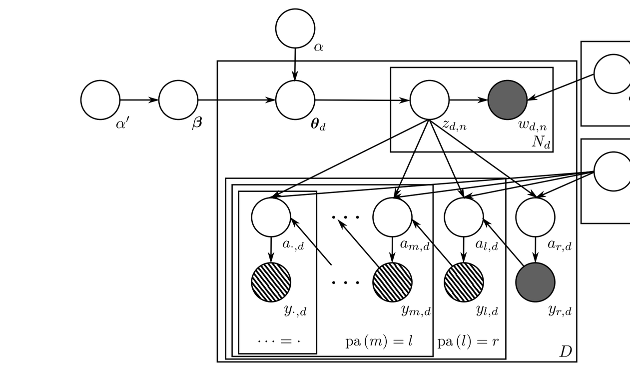


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [27]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [26]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right)(\alpha \boldsymbol{\beta}_k)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [28, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with two closely related models against the two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was given to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. Thus, we evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

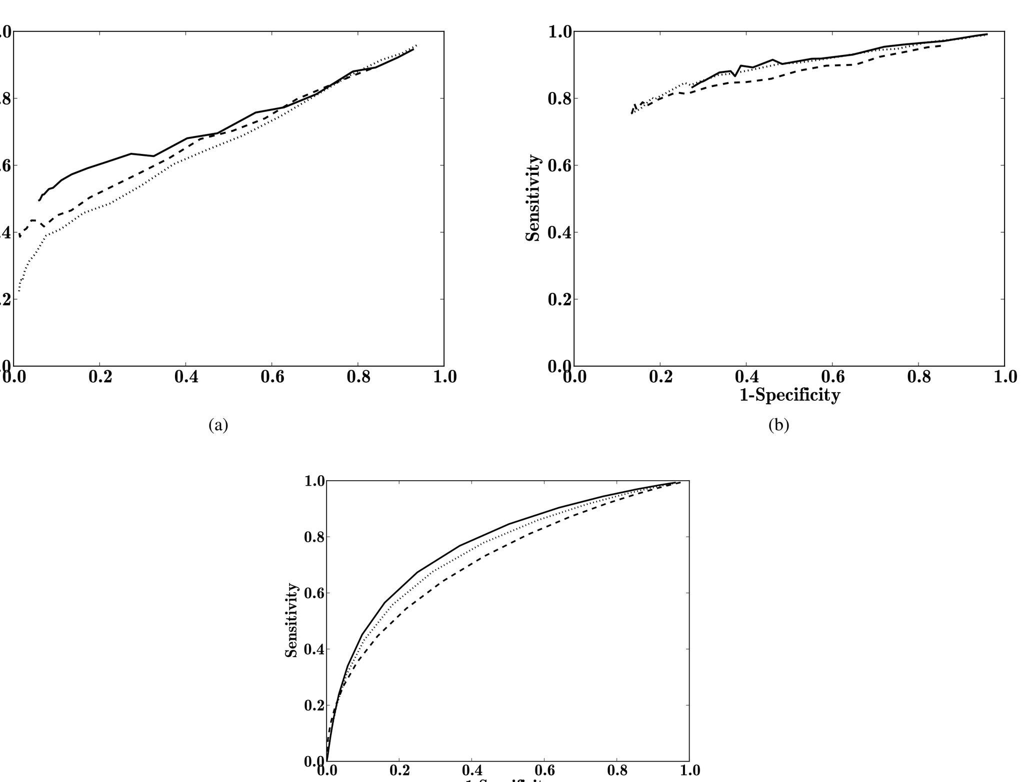


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics – sensitiviy (true positive rate) and 1-specificity (false positive rate).

The gold standard for comparison was derived from the testing set in each dataset. To make the comparison as fair as possible among models, ancestors of observed nodes in the label hierarchy were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive despite the fact that ancestors must also be positive. The gold standard defined in this way will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false

positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[27] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[28] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

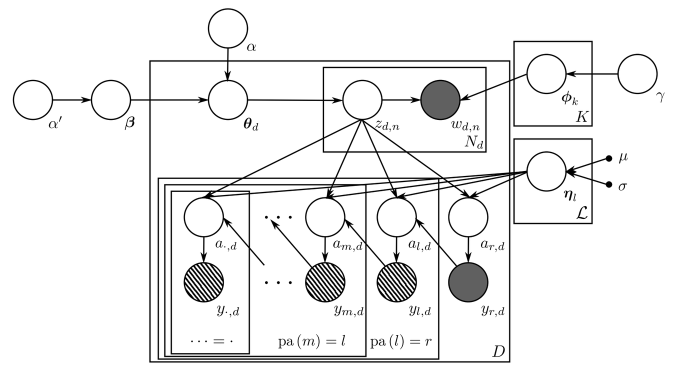


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \tag{1}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{2}$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \tag{3}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{4}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha \boldsymbol{\beta}_k)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha \boldsymbol{\beta}_k)^m \tag{5}$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [27, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with two closely related models against the two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the addition structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was given to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. Thus, we evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

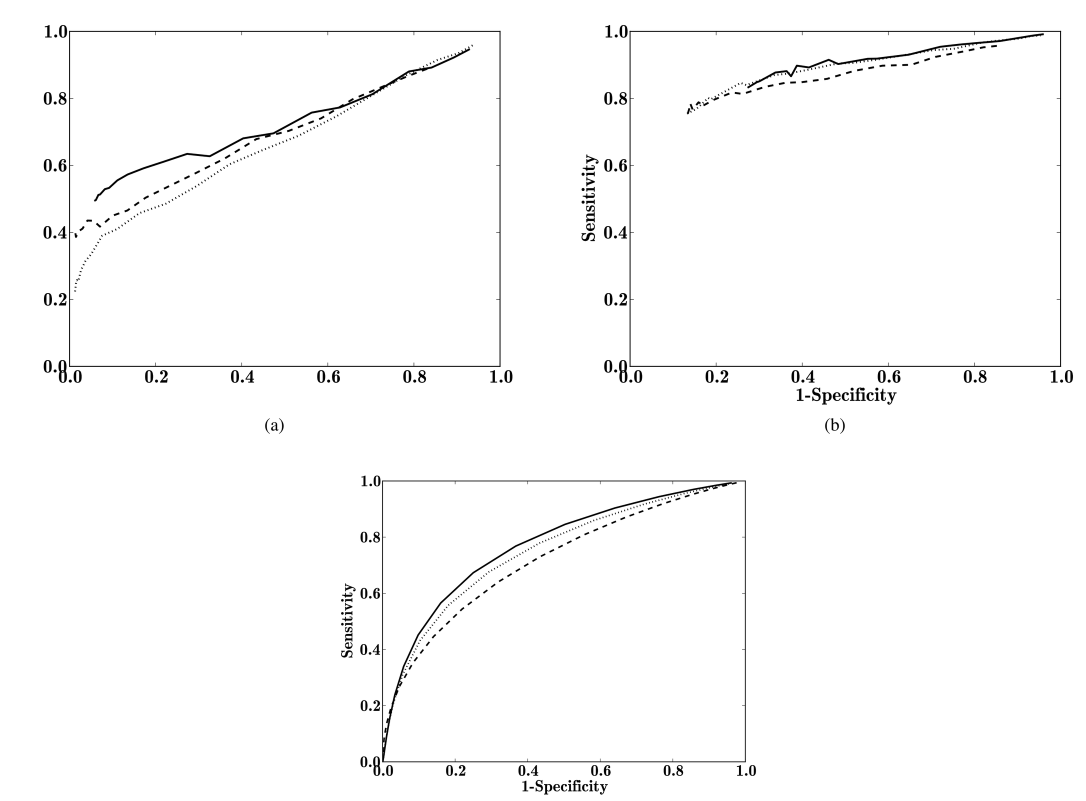


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics – sensitiviy (true positive rate) and 1-specificity (false positive rate).

The gold standard for comparison was derived from the testing set in each dataset. To make the comparison as fair as possible among models, ancestors of observed nodes in the label hierarchy were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive, despite the fact that, following the hierarchical constraints, ancestors must also be positive. Such a gold standard will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes, for instance, lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l \in \mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

The following are the performance results for the clinical data given a prior mean for the regression parameters of $\mu = -1.6$. The full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and a false

positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

The following are the predictive performance results for the retail product data given a prior mean for the regression parameters of $\mu = -2.2$. The full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off we examined sensitivity for a range of values for two different parameters - the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff.

For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference. However, as intended, they both highlight model performance under more or less stringent requirements for predicting positive labels.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirchlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. There are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other hierarchies. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic modeling and the hierarchical classification are carried out independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

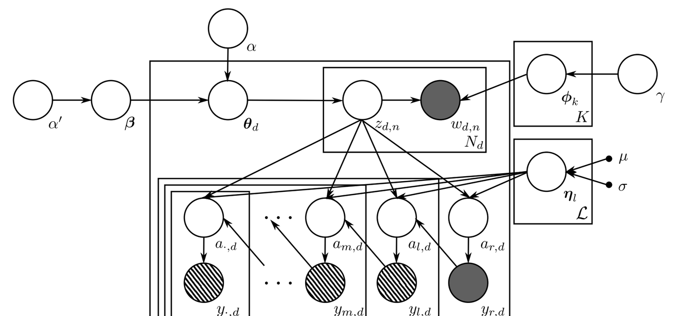


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma) \propto \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha\beta_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d})^2}{2}\right\} \quad (1)$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\Sigma}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\Sigma}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\Sigma}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}(a_{l,d} y_{l,d} > 0). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma(\alpha\beta_k)}{\Gamma\left(\alpha\beta_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) (\alpha\beta_k)^m \quad (5)$$

where $s(n,m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [27, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with two closely related models against the two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was given to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. Thus, we evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

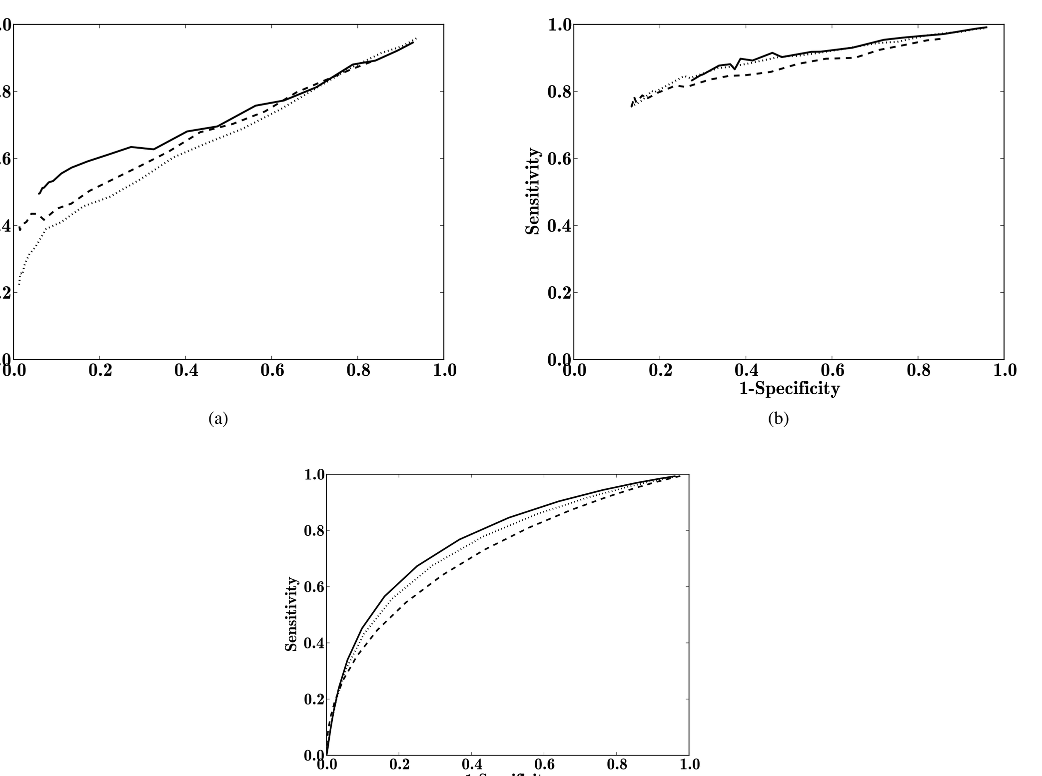


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics – sensitiviy (true positive rate) and 1-specificity (false positive rate).

The gold standard for comparison was derived from the testing set in each dataset. To make the comparison as fair as possible among models, ancestors of observed nodes in the label hierarchy were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive, despite the fact that, following the hierarchical constraints, ancestors must also be positive. Such a gold standard will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes, for instance, lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

Figure 2(a) shows the results for the clinical data. For instance, a prior mean for the regression of $\mu = -1.6$ yields the following performance: the full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and

a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

Figure 2(b) shows the results for the retail product data. For instance, a prior mean for the regression of $\mu = -2.2$ yields the following performance: the full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off, we examined sensitivity for a range of values for two different parameters – the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff. For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference and the auxiliary variable threshold is varied following inference.

Figure 2 shows the predictive performance of HSLDA relative to the three comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. http://www.dmoz.org/, 2002.

[3] Stanford network analysis platform. http://snap.stanford.edu/, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Hierarchically Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

1 Introduction

There exist many sources of unstructured data that have been partially or completely categorized by human editors. In this paper, we focus on unstructured text data that has been, at least in part, manually categorized. Examples include but are not limited to webpages and curated hierarchical directories of the same [2], product descriptions and catalogs, and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. In this work we show how to combine these two sources of information using a single model that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more. The models and techniques that we develop in this paper are applicable to other data as well, namely, any unstructured representations of data that have been hierarchically classified (e.g., image catalogs with bag-of-feature image representations).

There are several challenges entailed in incorporating a hierarchy of labels into the model. Among them, given a large set of potential labels (often thousands), each instance has only a small number of labels associated to it. Furthermore, there are no naturally occurring negative labeling in the data, and the absence of a label cannot always be interpreted as a negative labeling.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other structured label spaces. We extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic model and the hierarchical classification are inferred independently of each other.

The remainder of this paper is as follows. Section 2 introduces hierarchically supervised LDA (HSLDA), while Section 3 details a sampling approach to inference in HSLDA. Section 4 reviews related work, and Section 5 shows results from applying HSLDA to health care and web retail data.

2 Model

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked

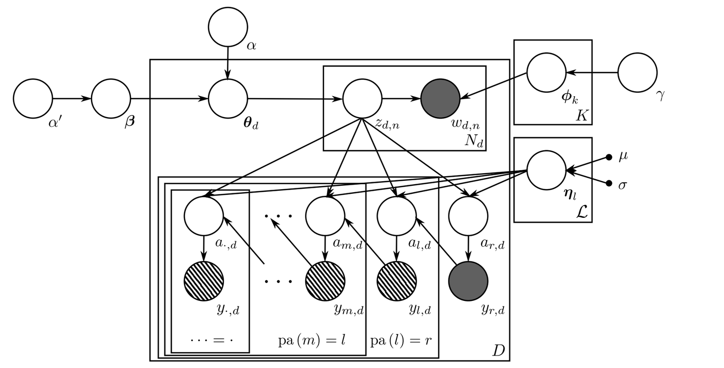


Figure 1: HSLDA graphical model

as not applying to a document then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The following procedure describes how to generate from the HSLDA generative model.

1. For each topic $k = 1, \ldots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$
2. For each label $l \in \mathcal{L}$
   - Draw a label application coefficient $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K(\alpha' \mathbf{1}_K)$
4. For each document $d = 1, \ldots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
   - For $n = 1, \ldots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$
   - Set $y_{r,d} = 1$
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$
     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

The second half of step 4 is a substantial part of our contribution to the general class of supervised LDA models. Here, each document is labeled generatively using a hierarchy of conditionally dependent probit regressors [14]. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e. $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Note that label $y_{l,d}$ can only be applied to document d if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but can be modified to accommodate different constraints). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this is that label predictors deeper in the label hierarchy are able to focus on finding specific, conditional labeling features. We believe this to be a significant source of the empirical label prediction improvement we observe experimentally. We test this hypothesis in Section 5.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (3) and the regression coefficients (2) which are analytic. This simplifies posterior inference substantially.

In the common case where no negative labels are observed (like the example applications we consider in Section 5), the model must be explicitly biased towards generating data that has negative labels in order to keep it from learning to assign all labels to all documents. This is a common problem in modeling unbalanced data. To see how this model can be biased in this way we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter above. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e. for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the ability of μ to bias out-of-sample label prediction performance in Section 5.

3 Inference

In this section we provide the conditional distributions required to draw samples from the HSLDA posterior distribution using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [16], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ in the following expressions. Let $\mathbf{a}$ be the set of all auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z}_d \backslash z_{n,d}$ the set $\mathbf{z}_d$ with element $z_{n,d}$ removed. The conditional posterior distribution of the latent topic indicators is

$$p\left(z_{n,d} = k \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \left(c_{(\cdot),d}^{k,-(n,d)} + \alpha\boldsymbol{\beta}_k\right) \frac{c_{w_{n,d},(\cdot)}^{k,-(n,d)} + \gamma}{\left(c_{(\cdot),(\cdot)}^{k,-(n,d)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \quad (1)$$

where $c_{v,d}^{k,-(n,d)}$ is the number of words of type v in document d assigned to topic k omitting the nth word of document d. The subscript $(\cdot)$'s indicate to sum over the range of the replaced variable, i.e. $c_{w_{n,d},(\cdot)}^{k,-(n,d)} = \sum_d c_{w_{n,d},d}^{k,-(n,d)}$. Here $\mathcal{L}_d$ is the set of labels which are observed for document d.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \quad (2)$$

where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}}\left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l\right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression [4]; the specific form of the update is a standard result from Bayesian normal linear regression [14]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [4].

$$p\left(a_{l,d,} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right). \quad (3)$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics [26]. Sampling $\boldsymbol{\beta}$ is done using the "direct assignment" method of Teh et al. [25]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \quad (4)$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + c_{(\cdot),d}^k\right)} s\left(c_{(\cdot),d}^k, m\right)\left(\alpha\boldsymbol{\beta}_k\right)^m \quad (5)$$

where $s(n, m)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

4 Related Work

In this work we extend supervised latent Dirichlet allocation (sLDA) [6] to take advantage of hierarchical supervision. sLDA is latent Dirichlet allocation (LDA) [7] augmented with per document "supervision," often taking the form of a single numerical or categorical label. It has been demonstrated that the signal provided by such supervision can result in better, task-specific document models and can also lead to good label prediction for out-of-sample data [6]. It also has been demonstrated that sLDA has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [6]. sLDA can be applied to data of the type we consider in this paper; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 5 we constrast HSLDA with sLDA applied in this way.

Other models that incorporate LDA and supervision include LabeledLDA[23] and DiscLDA[18]. Various applications of these models to computer vision and document networks have been explored [27, 9] . None of these models, however, leverage dependency structure in the label space.

In other work, researchers have classified documents into a hierarchy (a closely related task) with naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [21, 11, 17, 8].

5 Experiments

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

5.1 Data and Pre-Processing

5.1.1 Discharge Summaries and ICD-9 Codes

Discharge summaries are authored by clinicians to summarize patient hospitalization course. The summaries typically contain a record of patient complaints, findings and diagnoses, along with treatment and hospital course. For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] The ICD-9 codes are organized in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. It is worth noting that the coding can be noisy. Human coders sometimes disagree [1], tend to be more specific than sensitive in their assignments [5], and sometimes make mistakes [13].

The task of automatic ICD-9 coding has been investigated in the clinical domain. Methods range from manual rules to online learning [10, 15, 12]. Other work had leveraged larger datasets and experimented with K-nearest neighbor, Naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results [19, 24, 22, 20].

Our dataset was gathered from the clinical data warehouse of a large metropolitan hospital. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all the discharges from the hospital in 2009. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5,000 discharge summaries for training and 1,000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with highest document frequency (exclusive of names, places and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

5.1.2 Product Descriptions and Categorizations

Amazon.com, an online retail store, organizes its catalog of products in a mulitply-rooted hierarchy and provides textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

In this experiment, we obtained Amazon.com product categorization data from the Stanford Network Analysis Platform (SNAP) dataset [3]. Product descriptions were obtained separately from the Amazon.com website directly. We limited our dataset to the collection of DVDs in the product catalog.

Our dataset contains 15,130 product descriptions for training and 1,000 for testing. The product descriptions are shorter than the discharge summaries (91.89 terms on average, std dev=53.08). Overall, there are 2,691 unique codes. Products are assigned on average 9.01 codes (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

5.2 Comparison Models

We evaluated HSLDA along with two closely related models against the two datasets. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored), HSLDA fit by first performing LDA then fitting tree-conditional regressions. These models were chosen to highlight several aspects of HSLDA including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

sLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and sLDA is the additive structure imposed on the label space, a distinction that we hypothesized would result in a difference in predictive performance.

There are two components to HSLDA, LDA and a hierarchically constrained response. The second comparison model is HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space. In this comparison model, the separate inference processes do not allow the responses to influence the low dimensional structure inferred by LDA. Combined inference has been shown to improve performance in sLDA [6]. This comparison model examines not the structuring of the label space, but the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was given to the settings of the prior parameters for the regression coefficients. These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. Thus, we evaluated model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm
[2] http://www.nltk.org

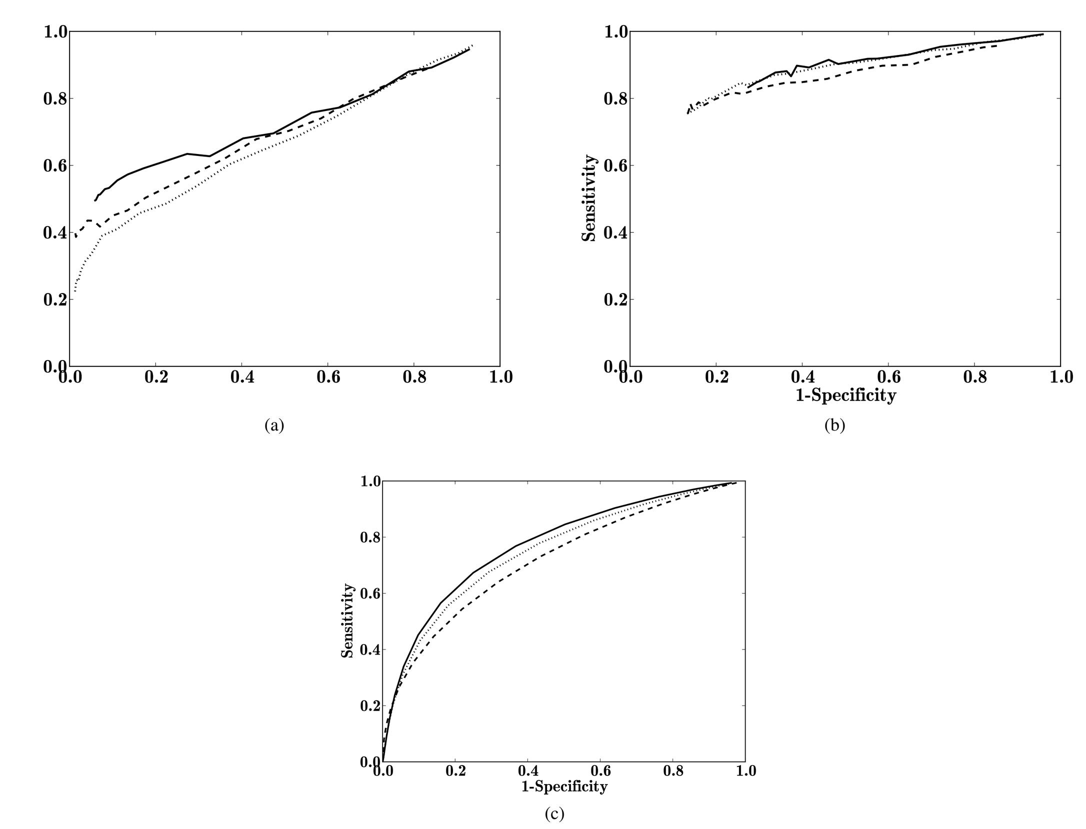


Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

5.3 Evaluation and Results

We evaluated our model, HSLDA, against the comparison models with a focus on predictive performance on held-out data. Prediction performance was measured with standard metrics – sensitiviy (true positive rate) and 1-specificity (false positive rate).

The gold standard for comparison was derived from the testing set in each dataset. To make the comparison as fair as possible among models, ancestors of observed nodes in the label hierarchy were ignored, observed nodes were considered positive and descendents of observed nodes were considered to be negative. In particular, since the sLDA model does not enforce the hierarchical constraints, its performance was found to be so poor on joint prediction that it was not included. The models can be compared on a more equal footing by considering only on the observed labels as being positive, despite the fact that, following the hierarchical constraints, ancestors must also be positive. Such a gold standard will likely inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes, for instance, lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives[5]. However, given that the label space is often large (as in our examples) it is a moderate assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by $p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$ for each test document, $\hat{d}$. For efficiency, the expectation of this probability distribution was estimated in the following way. Expectations of $\bar{\mathbf{z}}_d$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Using these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

Figure 2(a) shows the results for the clinical data. For instance, a prior mean for the regression of $\mu = -1.6$ yields the following performance: the full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the sLDA model had a true positive rate of 0.42 and

a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

Figure 2(b) shows the results for the retail product data. For instance, a prior mean for the regression of $\mu = -2.2$ yields the following performance: the full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the sLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data.

As sensitivity and specificity can always be traded off, we examined sensitivity for a range of values for two different parameters – the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff. For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference, and the auxiliary variable threshold is varied following inference.

Figure 2 shows the predictive performance of HSLDA relative to the two comparison models as a function of the prior mean on regression coefficients as a receiver operating characteristic (ROC) curve. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. For all values of the prior mean in both datasets, HSLDA outperforms sLDA with independent regressors. In the case of HSLDA with separately trained regression, HSLDA outperforms in the clinical dataset but performs equally well across the board with the retail product dataset.

6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of that kind of data possess structured labels, either hierarchically constrained or otherwise. That HSLDA directly addresses this kind of data is a large part of the motivation for this work. That it outperforms more straightforward approaches should be of interest to practitioners.

Variational Bayes has been the predominant estimation approach applied to sLDA models. Hierarchical probit regression makes for tractable Markov chain Monte Carlo SLDA inference, a benefit that should extend to other sLDA models should probit regression be used for response variable prediction there too.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

Extensions to this work include unbounded topic cardinality variants and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Utilizing different kinds of label structure is possible within this framework, but requires relaxing some of the simplifications we made in this paper for expositional purposes.

References

[1] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[2] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[3] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[4] J H Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

[5] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–5, 2005.

[6] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[8] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178, August 1998. ISSN 1066-8888.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. doi: 10.1214/09-AOAS309.

[10] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[12] R Farkas and G Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[13] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management*, 30:78–84, 2010.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[15] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[16] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[17] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[18] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[19] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

[20] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, 2008.

[21] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[22] S. Pakhomov, J. Buntrock, and C. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)*, 13(5):516–525, 2006.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[24] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

[26] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[27] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.