Hierarchically Supervised Latent Dirichlet Allocation

A. Perotte\textsuperscript{1}  N. Bartlett\textsuperscript{2}  N. Elhadad\textsuperscript{1}  F. Wood\textsuperscript{2}

\textsuperscript{1}\text{Department of Biomedical Informatics} \\
\text{Columbia University}

\textsuperscript{2}\text{Department of Statistics} \\
\text{Columbia University}
Introduction

- HSLDA: Hierarchically Supervised Latent Dirichlet Allocation
- Model of documents and labels
  - Structure in label space
- Large, real-world datasets
Introduction

- HSLDA: Hierarchically Supervised Latent Dirichlet Allocation
- Model of documents and labels
  - Structure in label space
- Large, real-world datasets
HSLDA: Hierarchically Supervised Latent Dirichlet Allocation

- Model of documents and labels
  - Structure in label space
- Large, real-world datasets
Amazon.com Data

Text: Product Descriptions: ~90 words/document
Labels: Product Categories: ~9 categories/document
Amazon.com Data

- Text: Product Descriptions: \( \sim 90 \) words/document
- Labels: Product Categories: \( \sim 9 \) categories/document
Clinical Data

- Text: Discharge summaries: \(~500\) words/document
- Labels: ICD9 codes: \(~8\) codes/document
Clinical Data

- Text: Discharge summaries: ~500 words/document
- Labels: ICD9 codes: ~8 codes/document
- Documents have latent structure
  - Points in low-dimensional space
- Latent dimensions
  - Distribution over words
- Regression parameters
  - Relationship between the latent space and the label space
Documents have latent structure
- Points in low-dimensional space

Latent dimensions
- Distribution over words

Regression parameters
- Relationship between the latent space and the label space

HSLDA
Documents have latent structure
- Points in low-dimensional space
- Latent dimensions
  - Distribution over words
- Regression parameters
  - Relationship between the latent space and the label space
Inference

- Collapsed Gibbs sampler
- Probit regression
  - Auxiliary variables allow for Gibbs sampling
Inference

- Collapsed Gibbs sampler
- Probit regression
  - Auxiliary variables allow for Gibbs sampling
Experiments

- Comparison models
  - sLDA with independent regressors
  - HSLDA fit by first performing LDA then fitting tree-conditional regressions

- Task: Prediction of out of sample labels
Experiments

- **Comparison models**
  - sLDA with independent regressors
  - HSLDA fit by first performing LDA then fitting tree-conditional regressions

- **Task**: Prediction of out of sample labels
## Example Topics

<table>
<thead>
<tr>
<th>Clinical Topics</th>
<th>Product Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASS</td>
<td>WOUND</td>
</tr>
<tr>
<td>CANCER</td>
<td>FOOT</td>
</tr>
<tr>
<td>RIGHT</td>
<td>CELLULITIS</td>
</tr>
<tr>
<td>BREAST</td>
<td>ULCER</td>
</tr>
<tr>
<td>CHEMOTHERAPY</td>
<td>LEFT</td>
</tr>
<tr>
<td>METASTATIC</td>
<td>ERYTHEMA</td>
</tr>
<tr>
<td>LEFT</td>
<td>PAIN</td>
</tr>
<tr>
<td>LYMPH</td>
<td>SWELLING</td>
</tr>
<tr>
<td>TUMOR</td>
<td>SKIN</td>
</tr>
<tr>
<td>BIOPSY</td>
<td>RIGHT</td>
</tr>
<tr>
<td>CARCINOMA</td>
<td>ABSCESS</td>
</tr>
<tr>
<td>LUNG</td>
<td>LEG</td>
</tr>
<tr>
<td>CHEMO</td>
<td>OSTEOMYELITIS</td>
</tr>
<tr>
<td>ADENOCARCINOMA</td>
<td>TOE</td>
</tr>
<tr>
<td>NODE</td>
<td>DRAINAGE</td>
</tr>
<tr>
<td>SERIES</td>
<td>BASEBALL</td>
</tr>
<tr>
<td>EPISODES</td>
<td>TEAM</td>
</tr>
<tr>
<td>SHOW</td>
<td>GAME</td>
</tr>
<tr>
<td>SEASON</td>
<td>PLAYERS</td>
</tr>
<tr>
<td>EPISODE</td>
<td>BASKETBALL</td>
</tr>
<tr>
<td>FIRST</td>
<td>SPORT</td>
</tr>
<tr>
<td>TELEVISION</td>
<td>SPORTS</td>
</tr>
<tr>
<td>SET</td>
<td>NEW</td>
</tr>
<tr>
<td>TIME</td>
<td>PLAYER</td>
</tr>
<tr>
<td>TWO</td>
<td>SEASON</td>
</tr>
<tr>
<td>SECOND</td>
<td>LEAGUE</td>
</tr>
<tr>
<td>ONE</td>
<td>FOOTBALL</td>
</tr>
<tr>
<td>CHARACTERS</td>
<td>STARS</td>
</tr>
<tr>
<td>DISC</td>
<td>FANS</td>
</tr>
<tr>
<td>GUEST</td>
<td>FIELD</td>
</tr>
</tbody>
</table>
Prediction

![Graph showing sensitivity and 1-specificity for different methods: HSLDA, sLDA, LDA + conditional regression.](image)
Summary

- HSLDA is a new topic model based on sLDA with hierarchical supervision.
- We derive an efficient Gibbs sampler for HSLDA.
- Label prediction can be improved with HSLDA if there exists significant structure in the label space.
Thank you!

- George Hripcsak, MD, MS
- National Library of Medicine