Approximate Inference Variational Inference

Task : eval. the post. distribution

$p(Z \mid X)$ — X observed; Z latent variables / params

could be $p(Z \mid X ; \theta)$

Often the case that

- space in which Z lives is very, eval'ing all possible (enumerating all) Z's
- posterior doesn't have a nice analytic form
  - * continuous var's: integrations might not be closed form
  - * discrete

alternative MCMC

Variational Inference (section 10.1 of PRML)

Applied to the Bayesian inference prob.

Z latent var's & params (set)

X set of observed vars

} might be N iid obs's

$X = \{x_1, \ldots, x_N\}$

$Z = \{z_1, \ldots, z_N\}$

Probabilistic model

$$p(X, Z)$$

Goal:

Find Posterior Dist $P(Z \mid X)$ and evidence $P(X)$

$$\ln p(X) = \mathcal{L}(q) + KL(q \| p)$$

where

$$\mathcal{L}(q) = \int q(Z) \cdot \ln \left\{ \frac{p(X, Z)}{q(Z)} \right\} dZ$$

$$KL(q \| p) = - \int q(Z) \ln \left\{ \frac{p(Z \mid X)}{q(Z)} \right\} dZ$$

multidimensional int's dimensionality

dim is # Z vars and dimensionality of each

$$\ln p(X) = \mathcal{L}(q) + KL(q \| p)$$

$$= \int q(z) \ln \left\{ \frac{P(x,z)}{q(z)} \right\} dz - \int q(z) \ln \left\{ \frac{P(z|x)}{q(z)} \right\} dz$$

KL → 0 when $q(z) = p(z|x)$

$$= \int q(z) \ln \left\{ \frac{P(z|x)P(x)}{q(z)} \right\} dz - \int q(z) \ln P(z|x) dz + \int q(z) \ln q(z) dz$$

$$= \int q(z) \ln P(z|x)\, dz + \int q(z) \ln p(x) \, dz - \int q(z) \ln q(z)\, dz$$

$$- \int q(z) \ln P(z|x)\, dz + \int q(z) \ln q(z)\, dz$$

$$= \int q(z) \ln p(x)\, dz = \ln p(x) \int q(z) dz = \ln p(x) \cdot 1$$

- Differs EM story only in that θ no longer appears

- Game: maximize $\mathcal{L}(q)$ (lower bound on the evidence) - equivalent to minimizing the KL div.

  If all q's possible then $q(z) = p(z|X)$ is minimum but $p(z|X)$ is complicated.

Approach

Restrict the family of distribution $q(z)$ to "simple" distributions, and then to seek the member of this family that most closely approx. $p(X)$

Note:

- choice of $q(z)$ all about tractability
- more complex $q(z)$'s are limited by computation
  - * no overfitting

Choices for $Q(Z)$

- Parameterized $q(Z|\omega)$ is governed by params ω
- Factorized $q(Z) = \prod_{i=1}^{M} q_i(Z_i)$] kind-of "independence" assumption

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i),$$

family of approximating dist's

no restrictions on the form of indiv. q_i's dists

Amongst all dist's in this family, which makes $\mathcal{L}(q)$ the largest?

$$\mathcal{L}(q) = \int q(Z) \ln \left\{ \frac{P(X,Z)}{q(Z)} \right\} dZ$$

use family $q(Z) = \prod_{i=1}^{M} q_i(Z_i)$ where Z_i is a subset of the latent vars

call $q_i(Z_i) \equiv q_i$

$$\mathcal{L}(q) = \int \left(\prod_i q_i\right) \ln \left\{ \frac{P(X,Z)}{\prod_i q_i} \right\} dZ$$

$q \in$ factorized family

EM-like objective, find conditions at optimal $\mathcal{L}(q)$ for each q_j.

Split out a single term q_j. Max $\mathcal{L}(q)$ w.r.t. a single term?

$$\mathcal{L}(q) = \int \left(\prod_i q_i\right)\left(\ln P(X,Z) - \sum_i \ln q_i\right) dZ$$

extract single factor q_j

$$= \underbrace{\int q_j \left(\prod_{i\neq j} q_i\right) \ln P(X,Z)\, dZ} - \underbrace{\int \left(\prod_i q_i\right)\left(\sum_i \ln q_i\right) dZ}$$

remember $q_j = q_j(z_j)$, $z_j \subseteq Z$

$$= \boxed{\int q_j \left\{ \int \cdots \int \prod_{i\neq j} q_i \ln P(X,Z)\, dz_1 \underset{\text{not } j}{\cdots} dz_M \right\} dz_j}$$

$$- \underbrace{\int q_j \left(\prod_{i\neq j} q_i\right)\left(\ln q_j + \sum_{i\neq j} \ln q_i\right) dZ}$$

operating on boxed quantity alone

$$= - \int q_j \underbrace{\left(\prod_{i\neq j} q_i\right)} \ln q_j \, dZ + \underbrace{\int q_j \left(\prod_{i\neq j} q_i\right) \cdot \left(\sum_{i\neq j} \ln q_i\right) dZ}$$

$$\boxed{= - \int q_j \ln q_j \, dz_j}$$

integrate to 1

const w.r.t. to q_j because $\int q_j \, dz_1 \left\{ \prod_{i\neq j} \quad dz_{\neq j} \right\} dz_j$

const w.r.t. q_j

Mess - clean it up

$$\mathcal{L}(q) = \int q_j \left\{ \int \cdots \int \prod_{i\neq j} q_i \ln P(X,Z)\, dz_1 \cdots dz_{j-1}\, dz_{j+1} \cdots dz_M \right\} dz_j$$

$$- \int q_j \ln q_j \, dz_j$$

If we define

$$\ln \tilde{p}(X, z_j) = \mathbb{E}_{i\neq j}\left[\ln p(X,Z)\right]$$

$$= \underbrace{\int \cdots \int}_{M-1} \ln P(X,Z)\, dz_1 \underset{\neq j}{\cdots} dz_M$$

$$\boxed{\mathcal{L}(q) = \int q_j \ln \tilde{p}(X, z_j)\, dz_j - \int q_j \ln q_j \, dz_j + \text{const}}$$

$$L(q) = \int q_j \ln \tilde{p}(X, Z_j)\, dZ_j - \int q_j \ln q_j \, dZ_j \quad + \text{const}$$

Goal: component-wise maximization of $L(q)$,
in particular, right now max $L(q)$ w.r.t. q_j

Recognize the $L(q)$ is a negative KL div.
between $\tilde{p}(X, Z_j)$ and $q_j(Z_j)$

Minimize KL divergence between $\tilde{p}(X, Z_j)$ and $q_j(Z_j)$.

The optimal $q_j^*(Z_j)$ is given by

forms of basis "all" variational approaches to inference

$$\boxed{\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const}}$$

- Condition of optimality

- Think about this for a second, the "coordinate / component" max is a function of Z_j (subset of params or lat. vars) - The right hand side is the joint dist of all obs & latent variables, but with all latent var's besides Z_j integrated out - this leaves a function of same var's Z_j on right hand side as well.

  - The specific dist. form of $q_j^*(Z_j)$ will often <u>emerge</u> from this rule.

    - $p(X, Z)$ probably has interesting cond. independencies to exploit.
    - in expectation (actually log) many / most of the terms will be absorbed into the constant but not all
    - coupling between approx. factors, say q_k and q_m $k \neq m$.
    - No closed form sol'n in general.

$$\ln q_j^*(z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const}$$

lacks normalization (const)

- only yields q_j^* up to a multiplicative factor
- one can normalize this distribution by either
- (usually)
  - inspection (will be become clear)
  - or by explicit normalization

$$q_j^*(z_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(x, z)])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(x, z)])\, dz_j}$$

- Set of the eqn's for all q_i, $i = 1..M$ is a set of "consistencies" conditions for the max. They are not an explicit sol'n, not in general closed form, and have to cycled through until numerical convergence.

State without proof Convergence of these interdependent updates is guaranteed because the objective is convex.

Teaching Example

Variational approximation to a full covariance Gaussian (2D)

Remember, in general $\vec{z} \in \mathbb{R}^2$ $\vec{z} \sim N(\mu, \Sigma)$

$$p(\vec{z}) \neq p(z_1)\, p(z_2) \quad \text{unless} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and $z_1 \perp\!\!\!\perp z_2$

Goal: find the independent Gaussian dist (diagonal Gaussian) that best approximates $p(\vec{z})$

Factorization $q(z_1)\, q(z_2)$