Variational Inference

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder

LECTURE 21

Variational Inference

- Inferring hidden variables
- More complicated models
- Connections to EM and Gibbs sampling
- Last HW

Setup

- $\vec{x} = x_{1:n}$ observations
- $\vec{z} = z_{1:m}$ hidden variables
- α fixed parameters
- Want the posterior distribution

$$p(z \mid x, \alpha) = \frac{p(z, x \mid \alpha)}{\int_z p(z, x \mid \alpha)}.$$

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \dots n$:
    2.1 Draw $z_i \sim \text{Mult}(\pi)$
    2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}$$

Main Idea

- We create a **variational distribution** over the latent variables

$$q(z_{1:m} \mid \nu) \tag{1}$$

- Find the settings of ν so that q is close to the posterior
- If $q == p$, then this is vanilla EM

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z \,|\, x)} \right] \tag{2}$$

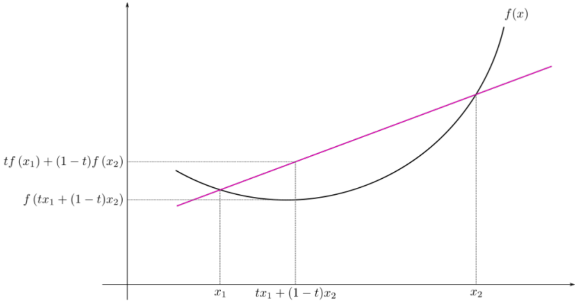
What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z \,|\, x)} \right] \tag{2}$$

- Characterizing KL divergence
  - If q and p are high, we're happy
  - If q is high but p isn't, we pay a praise
  - If q is low, we don't care
  - If $\mathrm{KL} = 0$, then distribution are equal

Concave Functions and Expectations



When f is concave

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

If you haven't seen this before, spend fifteen minutes to convince yourself that it's true

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z) \\
&= \log \int_z p(x,z)\frac{q(z)}{q(z)} \\
&= \log \left(\mathrm{E}_q \left[\frac{p(x,Z)}{q(z)} \right] \right) \\
&\geq \mathrm{E}_q[\log p(x,Z)] - \mathrm{E}_q[\log q(Z)]
\end{aligned}
$$

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z) \\
&= \log \int_z p(x,z) \frac{q(z)}{q(z)} \\
&= \log \left(\mathrm{E}_q \left[\frac{p(x,Z)}{q(z)} \right] \right) \\
&\geq \mathrm{E}_q[\log p(x,Z)] - \mathrm{E}_q[\log q(Z)]
\end{aligned}
$$

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{3}$$

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{3}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) || p(z \mid x)) &= \mathrm{E}_q \left[\log \frac{q(Z)}{p(Z \mid x)} \right] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z \mid x)] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z, x)] + \log p(x) \\
&= -(\mathrm{E}_q[\log p(Z, x)] - \mathrm{E}_q[\log q(Z)]) + \log p(x)
\end{aligned}$$

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{3}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) || p(z \mid x)) &= \mathrm{E}_q \left[\log \frac{q(Z)}{p(Z \mid x)} \right] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z \mid x)] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z, x)] + \log p(x) \\
&= -(\mathrm{E}_q[\log p(Z, x)] - \mathrm{E}_q[\log q(Z)]) + \log p(x)
\end{aligned}$$

- Negative of ELBO (plus constant); minimizing KL divergence is the same as maximizing ELBO

Mean field variational inference

- Assume that your variational distribution factorizes

$$q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j)$$

- You may want to group some hidden variables together
- Does not contain the true posterior because hidden variables are dependent

General Blueprint

- Choose q
- Derive ELBO
- Coordinate ascent of each q_i
- Repeat until convergence

Example: GMM

- Mean field family is

$$q(\mu_{1:K}, z_{1:n}) = \prod_k q(\mu_k \mid \tilde{\mu}_k, \tilde{\sigma}_k^2) \prod_i q(z_k \mid \phi_k) \tag{4}$$

- Induces the following ELBO

$$\left(\sum_{k=1}^{K} \mathrm{E}[\log p(\mu_k)] + \mathrm{H}(q(\mu_k)) \right) + \left(\sum_{i=1}^{n} \mathrm{E}[\log p(z_i)] + \mathrm{E}[\log p(x_i \mid z_i, \mu_{1:K})] + \mathrm{H}(q(z_i)) \right)$$

Expanding Expectations

- Expected log prior over mixture locations

Expanding Expectations

- Expected log prior over mixture locations

$$\mathrm{E}[\log p(\mu_k)] = -(1/2)\log 2\pi\sigma_0^2 - \mathrm{E}[\mu_k^2]/2\sigma_0^2 + \mathrm{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2$$

- Expected log prior over mixture assignments

Expanding Expectations

- Expected log prior over mixture locations

$$\mathrm{E}[\log p(\mu_k)] = -(1/2)\log 2\pi\sigma_0^2 - \mathrm{E}[\mu_k^2]/2\sigma_0^2 + \mathrm{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2$$

- Expected log prior over mixture assignments

$$\mathrm{E}[\log p(z_i)] = \log(1/K)$$

- Entropy over locations

Expanding Expectations

- Expected log prior over mixture locations

$$\mathrm{E}[\log p(\mu_k)] = -(1/2)\log 2\pi\sigma_0^2 - \mathrm{E}[\mu_k^2]/2\sigma_0^2 + \mathrm{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2$$

- Expected log prior over mixture assignments

$$\mathrm{E}[\log p(z_i)] = \log(1/K)$$

- Entropy over locations

$$\mathrm{H}(q(\mu_k)) = (1/2)\log 2\pi\tilde{\sigma}_k^2 + 1/2$$

- Entropy over assignments

Expanding Expectations

- Expected log prior over mixture locations

$$\mathrm{E}[\log p(\mu_k)] = -(1/2)\log 2\pi\sigma_0^2 - \mathrm{E}[\mu_k^2]/2\sigma_0^2 + \mathrm{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2$$

- Expected log prior over mixture assignments

$$\mathrm{E}[\log p(z_i)] = \log(1/K)$$

- Entropy over locations

$$\mathrm{H}(q(\mu_k)) = (1/2)\log 2\pi\tilde{\sigma}_k^2 + 1/2$$

- Entropy over assignments

$$\mathrm{H}(q(z_i)) = -\sum_{k=1}^{K}\phi_{ij}\log\phi_{ij}$$

Updates

- Update cluster assignments

$$q^*(z_i = k) \equiv \phi_{i,k} \propto \exp\left\{\log \pi_k + x_i \mathbb{E}_q\left[\mu\right] - \mathbb{E}_q\left[\mu_k^2\right]/2\right\} \tag{5}$$

- Update cluster centers

$$\tilde{\mu_k} = \frac{\mu_0/\sigma_0^2 + \sum_i \mathbb{E}_q\left[z_i^k\right] x_i}{1/\sigma_0^2 + \sum_i \mathbb{E}_q\left[z_i^k\right]} \tag{6}$$

Relationship with Gibbs Sampling

- Gibbs sampling: sample from the conditional distribution of all other variables
- Variational inference: each factor is set to the exponentiated log of the conditional
- Variational is easier to parallelize, Gibbs faster per step
- Gibbs typically easier to implement

In class

- Deriving variational inference for topic models
- Then you'll implement in your last homework