Variational Inference

References

These notes are based on the following papers:


Introduction

MCMC
- Theoretical guarantees (asymptotically) of sampling from the target density
- Computationally intensive but conceptually simple
- Handles multi-modal posterior distributions

Variational Inference
- No theoretical guarantees
- Good for big data and complex models
- Faster than MCMC but requires derivation of variational updates
- Can have problems with multi-modal posteriors

Introduction

- Variational methods based on calculus of variations
- Complex problem turned to a simpler one by decoupling degrees of freedom in the original problem
- Decoupling done by extending the original problem with additional variational parameters
Intuition

We first develop some intuition about variational methods using a simple example.

Write the log function as:
\[
\ln(x) = \min_{\lambda} \{\lambda x - \ln \lambda - 1\}
\]

Note:

• \(\lambda\) is the variational parameter
• For each value of \(x\), we need to compute the minimization of \(\lambda\).

Dashed lines correspond to:
\[
y = \lambda x - \ln \lambda - 1
\]

With different values of \(\lambda\)

• Varying \(\lambda\) produces a series of upper bounds:
\[
\ln(x) \leq \lambda x - \ln \lambda - 1
\]
• Minimizing \(\lambda\) produces the exact value for \(\ln(x)\)
• Note: \(\ln(x)\) is a concave function
Intuition

Why did we write \( \ln(x) = \min_{\lambda} \{ \lambda x - \ln\lambda - 1 \} \)?

• Comes from convex duality: a concave function \( f(x) \) can be represented by a dual function as
  \[
  f(x) = \min_{\lambda} \{ \lambda^T x - f^*(\lambda) \}
  \]
  Where
  \[
  f^*(\lambda) = \min_x \{ \lambda^T x - f(x) \}
  \]

• Applies to convex functions as well but you get a lower bound
Variational Inference

• Let $X$ be a set of observed variables (e.g. evidence variables)
• Let $Z$ be a set of latent variables
• Given inference query $P(Z|X) = \frac{P(X,Z)}{P(X)}$
• Need to compute $P(X) = \sum_Z P(X,Z)$ if $Z$ is discrete or $\int P(X,Z)dZ$ if continuous
• The denominator is typically very expensive to compute

The ELBO

• Goal: choose a density $q(z) \in Q$ which is the closest approximation to $p(z|x)$
• Here $Q$ is a family of densities over the latent variables
• Need to solve the following optimization problem:
$$q^*(z) = \arg\min_{q(z) \in Q} KL(q(z)||p(z|x))$$
The ELBO

\[ KL(q(z) || p(z|x)) \]

\[ = \int q(z) \log \frac{q(z)}{p(z|x)} dz = \int q(z) \log \frac{q(z)p(x)}{p(x,z)} dz \]

\[ = \int [q(z) \log q(z) + \log p(x) - \log p(x,z)] dz \]

\[ = \int q(z) \log q(z) dz + \int q(z) \log p(x) dz - \int q(z) \log p(x,z) dz \]

\[ = E_q(z)[\log q(z)] + E_q(z)[\log p(x)] - E_q(z)[\log p(x,z)] \]

\[ \text{Doesn't involve } q(x) \text{ so it can be taken out of the expectation} \]

Remember that this is very hard to compute because

\[ p(x) = \int p(x,z) dz \]

Rewrite as:

\[ \log p(x) = KL(q(z) || p(z|x)) + E_q(z)[\log p(x,z)] - E_q(z)[\log q(z)] \]

\[ = KL(q(z) || p(z|x)) + ELBO(q) \]

Where

\[ ELBO(q) = E_q(z)[\log p(x,z)] - E_q(z)[\log q(z)] \]
The ELBO

• Because KL divergence is $\geq 0$
  \[ \log p(x) = KL(q(z)||p(z)) + ELBO(q) \]
  \[ \Rightarrow \log p(x) \geq ELBO(q) \]

• $p(x)$ is the probability of the evidence, hence this is an evidence lower bound (ELBO)
• Instead of minimizing the KL divergence, we maximize the ELBO
  \[ ELBO(q) = E_{q(z)}[\log p(x, z)] - E_{q(z)}[\log q(z)] \]

Another point about $ELBO(q)$
\[
= E_{q(z)}[\log p(x, z)] - E_{q(z)}[\log q(z)] \\
= E_{q(z)}[\log p(x|z)] + E_{q(z)}[\log p(z)] - E_{q(z)}[\log q(z)] \\
= E_{q(z)}[\log p(x|z)] + \int q(z) \log p(z) dz - \int \log q(z) q(z) dz \\
= E_{q(z)}[\log p(x|z)] + \int q(z) \log \frac{p(z)}{q(z)} dz \\
= E_{q(z)}[\log p(x|z)] + KL(q(z)||p(z))
\]

This is an expected likelihood.
Places mass of $q(z)$ on configurations of the latent variables $z$ that explain the observed data $x$. This makes $q(z)$ resemble the prior $p(z)$.
Mean Field

In the mean-field variational family, the latent variables are:
- Mutually independent
- Each has its own factor (and parameters) in the variational family

\[ q(z) = \prod_{j=1}^{m} q_j(z_j) \]

Bayesian Mixture of Gaussians

- A Gaussian mixture model assumes there are \( K \) Gaussians that generate the data, each with its own mean \( \mu_k \) and variance \( \sigma_k \)

- We can make this a Bayesian model by putting a prior on the means of the \( K \) Gaussians
Bayesian Mixture of Gaussians

The Generative model:
• $\mu_k \sim N(0, \sigma^2)$ for $k = 1, \ldots, K$
• $c_i \sim \text{Categorical}(\frac{1}{K}, \ldots, \frac{1}{K})$ for $i = 1, \ldots, n$
• $x_i | c_i, \mu \sim N(c_i^T \mu, 1)$ for $i = 1, \ldots, n$

The joint density is:

$$p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i)p(x_i | c_i, \mu)$$
Bayesian Mixture of Gaussians

• Computing the evidence requires:
  \[ p(x) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i)p(x_i|c_i, \mu) \, d\mu \]
  
• This K-dimensional integral takes \( O(K^n) \) time to compute.
• Variational Inference to the rescue!

Bayesian Mixture of Gaussians

Joint density:
  \[ p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i)p(x_i|c_i, \mu) \]

Mean-field variational family:
  \[ q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \]

ELBO:
  \[ E_{q(\mu, c)}[\log p(\mu, c, x)] - E_{q(\mu, c)}[\log q(\mu, c)] \]
Bayesian Mixture of Gaussians

\[
ELBO(m, s^2, \varphi) = E_{q(\mu,c)}[\log p(\mu, c, x)] - E_{q(\mu, c)}[\log q(\mu, c)]
\]

\[
= E_{q(\mu,c)} \left[ \log \left( p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i|c_i, \mu) \right) \right] \\
- E_{q(\mu,c)} \left[ \log \left( \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \right) \right]
\]

\[
= \sum_{k=1}^{K} E_{q(\mu,c)} \left[ \log p(\mu_k); m_k, s_k^2 \right] \\
+ \sum_{i=1}^{n} \left( E_{q(\mu,c)} \left[ \log p(c_i); \varphi \right] + E_{q(\mu,c)} \left[ \log p(x_i; c_i, \mu); \varphi, m, s^2 \right] \right) \\
- \sum_{i=1}^{n} E_{q(\mu,c)} \left[ \log q(c_i; \varphi_i) \right] - \sum_{k=1}^{K} E_{q(\mu,c)} \left[ \log q(\mu_k; m_k, s_k^2) \right]
\]

With the ELBO, we now need to optimize the variational parameters.

One way to do this is coordinate ascent variational inference (CAVI) (Bishop 2006).

Works by optimizing each parameter while keeping the others fixed.

Need to come up with updates for \( \varphi_{ik}, m_k, s_k \).

Done iteratively until ELBO converges.
Bayesian Mixture of Gaussians

For CAVI:
• Uses the complete conditional of $z_j$ i.e. $p(z_j|z_{-j}, \mathbf{x})$
• Optimization uses the following:
  $$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j, z_{-j}, \mathbf{x})]\}$$
  This expectation is over all the other variational factors being fixed i.e. $\prod_{i \neq j} q_i(z_i)$
• We won’t go through the derivation. See (Bishop 2006) for details

Bayesian Mixture of Gaussians

For Bayesian Mixture of Gaussians:
1. Compute update for mixture assignments.
2. Compute update for mixture component means and variances.
1. Computing update for $\varphi_{lk}$

$$q^*(c_i; \varphi_i) \propto \exp\{\log p(c_i) + E[\log P(x_i|c_i, \mu); m, s^2]\}$$

$$\log p(c_i) = -1/K$$

$$\varphi_{lk} \propto \exp\{E[\mu_k; m_k, s_k^2]x_i - E[\mu_k^2; m_k, s_k^2]/2\}$$

(derivation left as an exercise)

Note: Expectation will be over $\prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{j \neq i} q(c_j; \varphi_j)$

2. Computing update for $m_k, s_k$

$$q(\mu_k; m_k, s_k^2) \propto \exp\left\{\log p(\mu_k) + \sum_{i=1}^{n} E[\log p(x_i|c_i, \mu); \varphi_i, m_{-k}, s_{-k}^2]\right\}$$

Note: Expectation will be over $\prod_{l \neq k} q(\mu_l; m_l, s_l^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$

This leads to update equations: (derivation left as an exercise)

$$m_k = \frac{\sum_{i} \varphi_{ik} x_i}{\sigma^2 + \sum_{i} \varphi_{ik}}$$

$$s_k^2 = \frac{1}{\sigma^2 + \sum_{i} \varphi_{ik}}$$
Concluding remarks

• It takes some work to derive variational inference equations
• Generic variational updates have been derived for special cases e.g. when complete conditional is in the exponential family