Variational Inference and Mean Field Method

Congyuan Duan

School of Mathematics, Sun Yat-sen University

December 29, 2020

Contents

Contents

Introduction

We are interested in the **posterior distribution**

$$p(z|x) \propto p(x|z)p(z)$$

However, we can't compute the posterior for many models. (Example:GMM)

Other methods: Integrated Nested Laplace Approximations(INLA), Monte Carlo Method...

Introduction

The basic idea of **Variational Inference** is, to pick an approximation $q(z)$ to the distribution from some tractable families, approximation should be as close as possible to the true posterior, $p^*(z) = p(z|x)$. This reduces inference to an optimization problem.

Contents

Kullback-Leibler Divergence

We measure the closeness of the two distributions with Kullback-Leibler (KL) divergence.
The KL divergence for variational inference is

$$KL(q||p) = E_q \left[\log \frac{q(z)}{p(z|x)} \right]$$

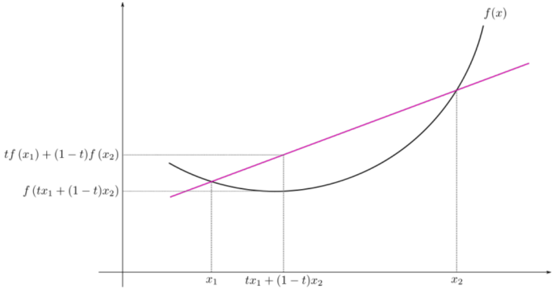
Note:reversing the arguments leads to a different kind of variational inference than we are discussing. In general, it's more computationally expensive than the algorithms we will study. We choose q so that we can take expectations.

The evidence lower bound

We actually can't minimize the KL divergence exactly, but we can minimize a function that is equal to it up to a constant. This is the **evidence lower bound (ELBO)**.

Recall Jensen's inequality as applied to probability distributions. When f is concave

$$f(E[X]) \geq E(f(X))$$



The evidence lower bound

We use Jensen's inequality on the log probability of the observations x

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

We define $ELBO(q) = E_q[\log \frac{p(x,z)}{q(z)}]$.

The evidence lower bound

First, note that

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

Now use the K-L divergence,

$$\begin{aligned}
KL(q(z)||p(z|x)) &= E_q\left[\log \frac{q(z)}{p(z|x)}\right] \\
&= E_q[\log q(z)] - E_q[\log p(z|x)] \\
&= E_q[\log q(z)] - E_q[\log p(z,x)] + \log p(x) \\
&= -(E_q[\log p(z,x)] - E_q[\log q(z)]) + \log p(x) \\
KL(q(z)||p(z|x)) + ELBO &= \log p(x)
\end{aligned}$$

This is the negative ELBO plus the log marginal probability of x. Notice that $\log p(x)$ does not depend on q. So, as a function of the variational distribution, minimizing the KL divergence is the same as maximizing the ELBO.

Contents

Mean field variational inference

In mean field variational inference, we assume that the variational family factorizes,

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

Each variable is independent.

Mean field variational inference

We now turn to optimizing the ELBO for this factorized distribution.
We will use **coordinate ascent inference**, interatively optimizing each variational distribution holding the others fixed.

Mean field variational inference

First, recall the chain rule and use it to decompose the joint,

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=i}^{m} p(z_j | z_{1:(j-1)}, x_{1:n})$$

Notice that the z variables can occur in any order in this chain.
The indexing from 1 to m is arbitrary.
Second, decompose the entropy of the variational distribution,

$$E[\log q(z_{1:m})] = \sum_{j=1}^{m} E_j[\log q(z_j)]$$

where E_j denotes an expectation with respect to $q(z_j)$.

Mean field variational inference

Third, define $\mathcal{L} := ELBO$, with these two facts, decompose $\mathcal{L}$,

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m} (E_j[\log p(z_j | z_{1:(j-1)}, x_{1:n})] - E_j[\log q(z_j)])$$

Employ the chain rule with the variable z_k as the last variable in the list.This leads to the objective function

$$\mathcal{L} = E[\log p(z_k | z_{-k}, x)] - E_j[\log q(z_k)] + \mathrm{const}$$

Write this objective as a function of $q(z_k)$:

$$\mathcal{L}_k = \int q(z_k) E_{-k}[\log p(z_k | z_{-k}, x)] dz_k - \int q(z_k) \log q(z_k) dz_k$$

Mean field variational inference

Take the derivative with respect to $q(z_k)$

$$\frac{d\mathcal{L}_k}{dq(z_k)} = E_{-k}[\log p(z_k|z_{-k}, x)] - \log q(z_k) - 1 = 0$$

This leads to the coordinate ascent update for $q(z_k)$

$$q^*(z_k) \propto \exp\{E_{-k}[\log p(z_k|z_{-k}, x)]\}$$

But the denominator of the posterior does not depend on z_j, so

$$q^*(z_k) \propto \exp\{E_{-k}[\log p(z_k, z_{-k}, x)]\}$$

The coordinate ascent algorithm is to iteratively update each $q(z_k)$. The ELBO converges to a local optimum. Use the resulting q is as a proxy for the true posterior

Exponential family conditionals

Suppose each conditional is in the exponential family

$$p(z_j|z_{-j}, x) = h(z_j)\exp\{\eta(z_{-j}, x)^T t(z_j) - a(\eta(z_{-j}, x))\}$$

This describes a lot of complicated models
-Bayesian mixtures of exponential families with conjugate priors
-Hierarchical HMMs
-Bayesian linear regression

Exponential family conditionals

Mean field variational inference is straightforward.
Compute the log of the conditional

$$\log p(z_j | z_{-j}, x) = \log h(z_j) + \eta(z_{-j}, x)^T t(z_j) - a(\eta(z_{-j}, x))$$

Compute the expectation with respect to $q(z_j)$

$$E[\log p(z_j | z_{-j}, x)] = \log h(z_j) + E[\eta(z_{-j}, x)]^T t(z_j) - E[a(\eta(z_{-j}, x))]$$

Noting that the last term does not depend on q_j, this means that

$$q^*(z_j) \propto h(z_j) \exp\{E[\eta(z_{-j}, x)]^T t(z_j)\}$$

So, the optimal $q(z_j)$ is in the same exponential family as the conditional.

Contents

Connection between VI and EM

In EM algorithm, we compute

$$\theta^{(n+1)} = \arg\max_{\theta} E_{z|x,\theta_n} \log p(x, z|\theta)$$

until it converges. So this gives rise to two steps. The E-step calculates the conditional expectation $E_{z|x,\theta_n} \log p(x, z|\theta)$, and the M-step maximizes the expectation.

Connection between VI and EM

When variational methods used in parameter estimation, notice that the ELBO is a function of the approximate distribution q and the unknown parameter θ,

$$ELBO(q, \theta) = \sum_z q(z|x) \log \frac{p(z, x|\theta)}{q(z|x)}$$

It is easy to prove that $p(z|x, \theta^{(k)}) = \arg\max_q ELBO(q, \theta^{(k)})$.

Connection between VI and EM

Procedure:

step1: $q^{(k+1)} = \arg\max_q ELBO(q, \theta^{(k)})$
step2: $\theta^{(k+1)} = \arg\max_\theta ELBO(q^{(k+1)}, \theta)$

Let's fix the approximate distribution $q(z|x)$ to be $p(z|x, \theta^{(k)})$.
Originally, we need to calculate $\theta^{(k+1)}$ as

$$\theta^{(k+1)} = \arg\max_\theta \sum \int p(z|x_i, \theta) \log p(x_i, z, \theta) dz - \int p(z|x_i, \theta) \log p(z|x_i, \theta) dz$$

However, we actually use

$$\theta^{(k+1)} = \arg\max_\theta \sum \int p(z|x_i, \theta^{(k)}) \log p(x_i, z, \theta) dz$$

Since after we set $\theta^{(k)}$ fixed, we can safely omit the terms in ELBO that don't contain θ.