Approximate Inference[2]

Mou Minghao

SSE, CUHK(SZ)

August 1, 2022

[2] the content mostly come from [Koller and Friedman, 2009]

Outline

- Introduction
  - Kullback-Leibler Divergence
  - Evidence Lower Bound (ELBO)
- Loopy Belief Propagation
- Mean-Field Approximation
  - Conditionally Conjugate Models
- Probabilistic Topic Models
  - Bags of Words Representation
  - Topic Model
- The Exponential Family: Conjugate Priors
  - Bernoulli Distribution and Beta Priors
  - Multinomial Distribution and Dirichlet Priors
  - Possion Distribution and Gamma Priors
  - Univariate Gaussian Distribution and Normal-Inverse-Gamma Priors
- Controlling Variance
  - Rao-Blackwellization
  - Control Variates
- References

Introduction

1. **Goal of Inference:**

   1.1 likelihood function of data

   1.2 marginal distribution

   1.3 conditional distribution

   1.4 modes of the density

2. **Approaches to Inference**

   2.1 exact inference algorithms

   2.1.1 brute-force

   2.1.2 variable elimination

   2.1.3 message passing (sum-product algorithm, belief propagation)

   2.1.4 junction tree algorithm

   2.2 approximate inference algorithms

   2.2.1 loopy belief algorithm

   2.2.2 variational inference + mean-field approximations

   2.2.3 MCMC / stochastical simulation / sampling

Kullback-Leibler Divergence

Definition (Kullback-Leibler Divergence)

$$KL(q||p) := \int_z q(z) \log \frac{q(z)}{p(z|x)} dz = \mathbb{E}_{z \sim q} \left[\log \frac{q(z)}{p(z|x)} \right] \tag{1}$$

Remark

1. *$KL(q||p)$ is nonnegative. (use Jensen's inequality)*
2. *$KL(\cdot||\cdot)$ is NOT a distance*

ELBO

To do variational Bayes, we want to minimize the KL divergence between our approximation q and our posterior p. However, we cant actually minimize this quantity (we will show why later), but we can minimize a function that is equal to it up to a constant. This function is known as the evidence lower bound (ELBO).

Theorem (Jensen's Inequality)

Assume f is convex, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{2}$$

The equality holds if and only if $X = \mathbb{E}(X)$

ELBO

Definition (ELBO)

The evidence lower bound (ELBO) of a probability model $p(x, z)$ and an approximation $q(z)$ is given by

$$ELBO(p, q) := \mathbb{E}_q[\log p(x, z)] - \mathbb{E}_q[\log q(z)] \tag{3}$$

This quantity is less than or equal to the evidence (log marginal probability of the observations), and We optimize this quantity (over densities $q(z)$) in Variational Bayes to find an "optimal approximation".

proof.

$$\log p(x) = \log \int_z p(x, z) dz = \log \int_z p(x, z) \frac{q(z)}{q(z)} dz \geq \int_z q(z) \log \frac{p(x, z)}{q(z)} dz$$

□

ELBO

Note that

1. We choose a family of variational distributions (i.e. a family of approximations) such that these two expectations can be computed.
2. The second expectation is the *entropy*, another quantity from information theory.
3. In variational inference, we find settings of the variational parameters ν that maximize the ELBO, which is equivalent to minimizing the KL divergence.

It is very easy to see that

$$KL(q||p) = \mathbb{E}_q\left[\frac{q(z)}{p(z|x)}\right] = -ELBO(p, q) + \log p(x) \tag{4}$$

We observe that this final line is the negative ELBO plus a constant (that does not depend on q). Therefore, we conclude that finding an approximation q that maximizes the ELBO is equivalent to finding the q that minimizes the KL divergence to the posterior.

Mean-Field Approximation

we assume the variational distribution over the latent variables factorizes as

$$q(z_1, z_2, ..., z_m) = \prod_{i=1}^{m} q_i(z_i) \tag{5}$$

We refer to $q_j(z_j)$, the variational approximation for a single latent variable, as a "local variational approximation".
we can also partition the latent variables into R groups $z_{G_1}, z_{G_2}, ..., z_{G_R}$, and use the approximation:

$$q(z_1, ..., z_m) = q(z_{G_1}, ..., z_{G_R}) = \prod_{i=1}^{R} q_{G_i}(z_{G_i}) \tag{6}$$

This is often called "generalized mean field" as compared to "naive mean field".
Typically, this approximation does not contain the true posterior (because the latent variables are dependent).

Mean-Field Approximation

We now want to optimize the ELBO in mean field variational inference. Typically, coordinate ascent optimization is used. We have to note that this is not the only way to optimize the ELBO in mean field approximation. One can also do gradient ascent.

First, note that by the chain rule

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j | z_{1:j}, x_{1:n}) \tag{7}$$

Note that the latent variables in this product can occur in any order, then

$$\mathbb{E}_q[\log(q)] = \sum_{i=1}^{m} \mathbb{E}_{q_i}[\log(q_i(z_i))] \tag{8}$$

Mean-Field Approximation

Use (7) and (8), we can decompose $ELBO(p, q)$ into a nice form

$$\begin{aligned} ELBO(p,q) =& \mathbb{E}_q\left[p(x,z)\right] - \mathbb{E}_q[\log q(z)] \\ =& \log p(x_{1:n}) + \sum_{i=1}^{m} \left[\mathbb{E}_q[\log p(z_i|z_{1:(i-1)}, x_{1:n})] - \mathbb{E}_{q_i}[\log q_i(z_i)]\right] \end{aligned} \tag{9}$$

Next, we want to maximize $ELBO(p, q)$

$$\begin{aligned} &\max_{q_j} ELBO(p,q) \\ &= \max_{q_j} \mathbb{E}_q[\log p(z_j|z_{-j}, x)] - \mathbb{E}_{q_j}[\log q(z_j)] \\ &= \max_{q_j} \left[\int_{z_j} q_j(z_j)\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]dz_j - \int_{z_j} q(z_j)\log q(z_j)dz_j\right] \end{aligned} \tag{10}$$

Mean-Field Approximation

use Lagrange multipliers, and set the derivative to 0, we have

$$q_j^*(z_j) \propto \exp[\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]] \propto \exp[\mathbb{E}_{q_{-j}}[\log p(z, x)]] \tag{11}$$

This coordinate ascent procedure convergences to a local maximum. The coordinate ascent update for $q_j(z_j)$ only depends on the other.

Remark (Summary)

1. *we first defined a family of approximations called mean field approximations, in which there are no dependencies between latent variables*
2. *Then we decomposed the ELBO into a nice form under mean field assumptions*
3. *we derived a coordinate ascent updates to iteratively optimize each local variational approximation under mean field assumptions*

Complete Conditionals

For the large class of conditionally conjugate models, we can easily perform optimization with a coordinate-ascent algorithm, one in which we iteratively optimize each varational parameter while holding the others fixed

Definition (Complete Conditional)

A complete conditional is the conditional distribution of a latent variable given the observations and the other latent variables in the model

Dirichlet Distribution

Definition (Dirichlet Distribution)

θ is said to follow Dirichlet distribution (i.e. $\theta \sim Dir(\alpha)$) if its density is

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_n^{\alpha_n - 1} \tag{12}$$

Rao-Blackwellization

Rao-Blackwellization reduces the variance of a random variable by replacing it with its conditional expectation with respect to a subset of the variables

Rao-Blackwellization replaces a function of two variables with its conditional expectation. Consider two random variables X and Y, and a function $J(X, Y)$. Our goal is to compute its expectation $\mathbb{E}[J(X, Y)]$ w.r.t. the joint distribution of X and Y

Define $\hat{J}(X) := \mathbb{E}[J(X, Y)|X]$ and note that $\mathbb{E}[J(X)] = \mathbb{E}[J(X, Y)]$. This means that $\hat{J}(X)$ can be used in place of $J(X, Y)$ in a Monte Carlo approximation of $J(X, Y)$, also

$$Var(\hat{J}(X)) = Var(J(X, Y)) - \mathbb{E}[(J(X, Y) - \hat{J}(X))^2] \qquad (13)$$

Thus, $\hat{J}(X)$ has a lower variance than $J(X, Y)$

Control Variates

A control variate is a family of functions with equivalent expectation. Consider a function h, which has a finite first moment, and a scalar a. Define $\hat{f}$ as

$$\hat{f}(z) := f(z) - a(h(z) - \mathbb{E}[h(z)]) \tag{14}$$

This is a family of functions, indexed by a, and note that $\mathbb{E}_q[\hat{f}] = \mathbb{E}_q[f]$ as required. Given a particular function h, we can choose a to minimize the variance of $\hat{f}$

First note that

$$Var(\hat{f}) = Var(f) + a^2 Var(h) - 2aCov(f, h) \tag{15}$$

This equation implies that good control variates have high covariance with the function whose expectation is being computed. It is easy to show

$$a^* = \frac{Cov(f, h)}{Var(h)} \tag{16}$$

References I

Koller, D. and Friedman, N. (2009).
Probabilistic graphical models: principles and techniques.
MIT press.