Machine Learning **Summer 2017**

Topic 1: Variational Inference

Author: Andy Lee

1.1 Introduction

In general, variational inference techniques are used to approximate difficult-to-compute probability densities such as intractable posterior densities in bayesian inference. These notes serve largely to supplement David Blei's tutorial *Variational Inference: A Review for Statisticians* [1]. In addition, many of my derivations are drawn from Eric Jang's wonderful tutorial on Variational Inference methods [2].

1.1.1 Background

Consider the general problem where we have a set of latent variables $z = \{z_1, \cdots, z_m\}$ and observations $x = \{x_1, \cdots x_n\}$. Recall in the bayesian framework, we have the following quantities of interest.

- **prior** $p(z)$ - prior density over latent variables
- **likelihood** $p(x|z)$ - likelihood of data over latent variables
- **posterior** $p(z|x)$ - posterior (how well latent variables describe data)

We are interested in maximizing the posterior to derive the MAP estimate z^* in order to perform inference

$$p(x_{\text{new}}|z^*)$$

To set up our inference problem, we can note

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

Note $p(x)$ is typically known as the **evidence** and is often intractable to compute. Moreover, notice that we cannot compute a closed-form solution to $p(z|x)$ without $p(x)$.

1.1.2 Variational Inference Approach

The typical approach is to use sampling techniques like MCMC to get an approximation to the otherwise intractable quantity $p(x)$. However, there are problems for which this approach will not work well, in particular when datasets are large or models are very complex. The variational inference takes a different approach to sampling.

Rather than use sampling, variational inference turns to optimization. First, we posit a family of approximate densities $\mathscr{D}$ over the space of latent variables z. We will try to find $q \in \mathscr{D}$ that minimizes the Kullback-Leibler divergence. Mathematically, we have the following optimization problem

$$q^*(z) = \underset{q(z|x)\in\mathscr{D}}{\arg\min}\, \text{KL}\Big(q(z|x)||p(z|x)\Big) \tag{1.1}$$

1.2 Deriving Variational Bound

In this section, we derive the **variational bound** (also known as **evidence lower bound**), a quantity that is used to approximate the evidence. Note that in this section, we use q and q_ϕ, where the latter explicitly denotes that the density is parametrized by some parameter ϕ on which we are optimizing over.

Recall, we are interested in solving the following optimization problem.

$$\begin{aligned}
q^*(z) &= \operatorname*{arg\,min}_{q_\phi(z|x)\in\mathscr{D}} \Big(\mathrm{KL}(q_\phi(z|x)||p(z|x))\Big) \\
&= \operatorname*{arg\,min}_{q_\phi(z|x)\in\mathscr{D}} \Big(\log p(x) + \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z,x)]\Big) \\
&= \operatorname*{arg\,min}_{q_\phi(z|x)\in\mathscr{D}} \Big(\mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z,x)]\Big) \quad (p(x) \text{ is a constant with respect to } q) \\
&= \operatorname*{arg\,max}_{q_\phi(z|x)\in\mathscr{D}} \Big(\mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z,x)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)]\Big) \\
&= \operatorname*{arg\,max}_{q_\phi(z|x)\in\mathscr{D}} \Big(\mathcal{L}(q_\phi)\Big)
\end{aligned}$$

The function $\mathcal{L}$ defined above is called the *evidence lower bound* or *variational bound.*

1.2.1 Intuition for variational bound

Note that we can rewrite the variational bound in a more intuitive form.

$$\begin{aligned}
\mathcal{L}(q_\phi) &= \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z,x)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)] \\
&= \mathop{\mathbb{E}}_{z\sim q_\phi}\Big[\log p(x|z) + \log p(z)\Big] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)] \\
&= \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(x|z)] + \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)] \\
&= \mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(x|z)] - \mathrm{KL}(q(z|x)||p(z))
\end{aligned}$$

In this form, we can see that the variational bound is the expected log-likelihood of the data and the KL divergence between the prior $p(z)$ and $q(z|x)$. By maximizing the evidence lower bound, the first term will encourage densities q such that the latent variables explain the observed data. The second term will encourage the variational density q to be close to the prior $p(z)$.

1.2.2 Why is $\mathcal{L}$ called the evidence lower bound?

$$\begin{aligned}
\mathrm{KL}\Big(q(z|x)||p(z|x)\Big) &= \log p(x) - \big[\mathop{\mathbb{E}}_{z\sim q_\phi}[\log p(z,x)] - \mathop{\mathbb{E}}_{z\sim q_\phi}[\log q_\phi(z|x)]\big] \\
&= \log p(x) - \mathcal{L}(q)
\end{aligned}$$

The next equation gives intuition for the naming.

$$\log p(x) = \mathcal{L}(q_\phi) + \mathrm{KL}\Bigg(q_\phi(z|x)||p(z|x)\Bigg) \tag{1.2}$$

Since the KL-divergence is always non-negative, it follows that the variational bound is a lower bound on the evidence $p(x)$.

1.3 Supplementary

In this section, we provide derivations of quantities used to arrive at the variational lower bound.

Lemma 1.1

$$KL(Q_\phi(z|x)||P(z|x)) = \underset{z\sim q_\phi}{\mathbb{E}}[\log q_\phi(z|x)] - \underset{z\sim q_\phi}{\mathbb{E}}[\log p(z,x)] + \log p(x)$$

Proof:

$$\begin{aligned}
KL(Q_\phi(z|x)||P(z|x)) &= \sum_{z\in Z} q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z|x)} \\
&= \sum_{z\in Z} q_\phi(z|x) \log \frac{q_\phi(z|x)p(x)}{p(z,x)} \text{ (since } p(z|x) = \frac{p(z,x)}{p(x)} \text{)} \\
&= \sum_{z\in Z} q_\phi(z|x) \big(\log \frac{q_\phi(z|x)}{p(z,x)} + \log p(x) \big) \\
&= \Big(\sum_{z} q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z,x)} \Big) + \Big(\sum_{z} \log p(x) q_\phi(z|x) \Big) \\
&= \Big(\sum_{z} q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z,x)} \Big) + \Big(\log p(x) \sum_{z} q_\phi(z|x) \Big) \\
&= \log p(x) + \Big(\sum_{z} q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z,x)} \Big) \\
&= \log p(x) + \underset{z\sim q_\phi}{\mathbb{E}}[\log q_\phi(z|x)] - \underset{z\sim q_\phi}{\mathbb{E}}[\log p(z,x)]
\end{aligned}$$

■

References

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. 2016.

[2] Eric Jang. A beginner's guide to variational method: Mean-field approximation, August 2016.