A short note on the variational inference

Yen-Chi Chen
University of Washington
April 13, 2020

Variational inference (VI; also known as variational approximation) is a popular tool in machine learning. It has became more and more popular in statistics communities as well. In short, VI is a method to approximate an intractable quantity using a tractable quantity. It can be used in both Frequentist estimation as well as Bayesian inference.

However, the fact that VI can be used in both Frequentist and Bayesian inference had made VI sometimes confusing. So here I will try to explain how VI can be used in both cases and how the two problems are associated.

1 Approximating an MLE (Frequentist)

Consider a regular latent variables problem where we observe IID

$$X_1, \cdots, X_n \sim p$$

and each observation has a latent variable Z that is unobserved. Namely, the complete data should be

$$(X_1, Z_1), \cdots, (X_n, Z_n)$$

but we only observe $X_1, \cdots, X_n$.

In the latent variable problem, we often place a parametric model on the complete-data distribution:

$$p(x, z; \lambda), \quad \lambda \in \Lambda.$$

This parametric model implies the observed model

$$p(x; \lambda) = \int p(x, z; \lambda) dz$$

that generates our observations.

In this case, the MLE is defined as

$$\widehat{\lambda} = \mathsf{argmax}_\lambda \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \lambda).$$

Namely, we maximizes the *observed log-likelihood* $\ell(\lambda|x) = p(x; \lambda)$. A population version of this problem is

$$\lambda^* = \mathsf{argmax}_\lambda \mathbb{E}(\log p(X_1; \lambda)).$$

In most cases, the MLE does not have a closed-form so we need to use numerical procedure such as the EM algorithm to compute it. However, EM algorithm may not have a simple form and can still be intractable (this occurs a lot when the Q function in the EM algorithm is complicated).

The VI offers a remedy to this problem. Note that we can write the observed model as

$$\begin{aligned} p(x;\lambda) &= \int p(x,z;\lambda)dz \\ &= \int \frac{p(x,z;\lambda)}{q(z;\omega)} q(z;\omega)dz \\ &= \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\frac{p(x,Z;\lambda)}{q(Z;\omega)}\right), \end{aligned}$$

where $\omega \in \Omega$ is another class of parameters and $Q = \{q(\cdot;\omega) : \omega \in \Omega\}$ is called the variational family. The variational family often consists of parametric densities that are easy to compute and sample.

Using the Jensen's inequality, the observed log-likelihood function becomes

$$\begin{aligned} \ell(\lambda|x) &= \log p(x;\lambda) \\ &= \log \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\frac{p(x,Z;\lambda)}{q(Z;\omega)}\right) \\ &\geq \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\log \frac{p(x,Z;\lambda)}{q(Z;\omega)}\right) \\ &= \mathbb{E}_{Z\sim q(\cdot;\omega)} \log p(x,Z;\lambda) - \mathbb{E}_{Z\sim q(\cdot;\omega)}(\log q(Z;\omega)) \\ &= \mathsf{ELBO}(\omega,\lambda|x). \end{aligned}$$

The quantity $\mathsf{ELBO}(\omega,\lambda|x)$ is called the *evidence lower bound*. Note that sometime people would write it as $\mathsf{ELBO}(q)$ and abbreviate λ, x but here for completeness I would keep both of them. When we observe n observations, we write

$$\mathsf{ELBO}_n(\omega,\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{ELBO}(\omega,\lambda|X_i).$$

With the ELBO, the idea of variational inference is very simple. Instead of maximizing the intractable log-likelihood, we maximize ELBO to find our estimator. Namely, our estimator is

$$(\widehat{\omega}_{\mathsf{VI}}, \widehat{\lambda}_{\mathsf{VI}}) = \mathsf{argmax}_{\omega,\lambda}\mathsf{ELBO}_n(\omega,\lambda).$$

Again, in general there is no closed-form to the above estimators. But we can solve this maximization problem numerically. The power of VI is that the variational family is designed so that evaluation and computation are easy. So the expectation $\mathbb{E}_{Z\sim q(\cdot;\omega)}\cdots)$ that appears in the ELBO is a tractable quantity.

A famous method called *mean field variational family* is the collection of densities

$$q(z;\omega) = \prod_{j=1}^{d} q(z_j;\omega_j),$$

where $z = (z_1,\cdots,z_d)$. In this case, sampling of z can be decomposed into sampling each coordinate independently and the optimization of ω_j can be obtained by the *coordinate ascent variational inference* algorithm in the following book:

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer New York.

One thing to keep in mind is that when using VI, we are no longer solving the original problem. In general, the VI estimator will not converge to the MLE (in the Frequentist setting) but instead, it will still converge to a population quantity; see the following paper:

Chen, Y. C., Wang, Y. S., & Erosheva, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. The Annals of Applied Statistics, 12(2), 846-876.

2 Density evaluation problem (Bayesian)

A good review from the Bayesian perspective is

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518), 859-877.

Consider a simple Bayesian setting where

$$X|\theta \sim p(x|\theta), \quad \theta \sim \pi(\theta).$$

Here, X is the random variable for the observation and θ is the underlying parameter of the distribution and π is the prior. The Bayesian inference relies heavily on the posterior distribution (density):

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \propto p(x|\theta)\pi(\theta) = p(x,\theta).$$

Here we write $p(x,\theta) = p(x|\theta)\pi(\theta)$ for simplicity. In addition to the posterior distribution, the evidence $p(x) = \int p(x,\theta)d\theta$ is also sometimes of interest.

Since $p(x,\theta) = p(x|\theta)\pi(\theta)$, it is often very easy to evaluate the value of $p(x,\theta)$ for any given x,θ. Although the joint distribution is easy to compute, both the posterior and the evidence are often intractable. Take the evidence as an example, it can be written as the integral $p(x) = \int p(x,\theta)d\theta$. When the dimension of θ is large (say a mixture model), this integration is very difficult to evaluate even if we can easily compute $p(x,\theta)$.

Here is an interesting note. The problem of evaluating the evidence and the problem of evaluating the posterior distribution are the same via the following relation:

$$\pi(\theta|x) = \frac{p(x,\theta)}{p(x)}.$$

$p(x,\theta)$ is tractable so we can easily convert the evidence and the posterior to each other.

To obtain a tractable approximation of $\pi(\theta|x)$, we use the idea of VI. The quantity θ now plays the role of latent variable z in the previous section. Let $q(\theta;\omega)$ be a computable density and

$$Q = \{q(\cdot;\omega) : \omega \in \Omega\}$$

be the variational family. We attempt to find the best density $q(;\omega^*) \in Q$ such that

$$q(\cdot;\omega^*) = \mathsf{argmin}_{q\in Q}\mathsf{KL}(q(\cdot)||\pi(\cdot|x)),$$

where KL is the Kullback-Leibler divergence. Namely,

$$\omega^* = \mathsf{argmin}_{\omega}\mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)),$$

However, minimizing the KL divergence may again ran into the same computation problem (need to evaluate the integral). The idea of VI uses the following insight:

$$\begin{aligned}\log p(x) &= \mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)) + \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log p(x,\theta)) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega)) \\ &= \mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)) + \mathsf{ELBO}(\omega|x).\end{aligned}$$

Therefore, minimizing the KL divergence is equivalent to maximizing ELBO.

The variational approximation chooses

$$\omega^*_{\mathsf{VI}} = \mathsf{argmax}_{\omega}\mathsf{ELBO}(\omega|x)$$

and use $q(\theta;\omega^*_{VI})$ as an approximation to $\pi(\theta|x)$.

In the case of observing $X_1,\cdots,X_n$, the ELBO will be

$$\begin{aligned}\mathsf{ELBO}_n(\omega) &= \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log p(X_1,\cdots,X_n,\theta)) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega)) \\ &= \mathbb{E}_{\theta\sim q(\cdot;\omega)}\left(\log\pi(\theta) + \sum_{i=1}^n \log p(X_i|\theta)\right) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega))\end{aligned}$$

and we estimate ω using

$$\widehat{\omega}_{\mathsf{VI}} = \mathsf{argmax}_{\omega}\mathsf{ELBO}_n(\omega).$$

Again, this maximization is often tractable since the expectation is with respect to the variational distribution q, which is by design easy to compute.

The posterior is then approximated by

$$\pi(\theta|X_1,\cdots,X_n) \approx p(\theta;\widehat{\omega}_{\mathsf{VI}})$$

and the evidence is approximated by

$$p(X_1,\cdots,X_n) \approx \frac{\pi(\theta)\prod_{i=1}^n p(X_1|\theta)}{p(\theta;\widehat{\omega}_{\mathsf{VI}})}.$$

Note that the approximated evidence may depend on θ because it is an approximation rather than an exact value.