Variational inference

Variational inference is a high-level paradigm for estimating a posterior distribution when computing it explicitly is intractable. Unlike expectation maximization, variational inference estimates a closed form density function for the posterior rather than a point estimate for the latent variables. Thus, variational inference is also different from MCMC methods in that it involves approximating the posterior with an analytical approximation rather than via sampling.

More specifically, variational inference is used in situations in which we have a model that involves hidden random variables Z, observed data X, and joint model over the hidden and observed data $p(z, x)$. Our goal is to perform inference on the hidden data via the posterior distribution over Z as given by Bayes theorem:

$$p(z \mid x) := \frac{p(x \mid z)p(z)}{p(x)}$$

Variational inference casts the problem of computing the posterior as that of finding another distribution $q(z)$ that is "close" to $p(z \mid x)$. Ideally, $q(z)$ is easier to evaluate than $p(z \mid x)$, and, if $p(z \mid x)$ and $q(z)$ are similar, then we can use $q(z)$ as a replacement for $p(z \mid x)$ in our inference tasks.

We restrict our search for $q(z)$ to a family of surrogate distributions over Z, called the **variational distribution family**, denoted $q(z \mid \phi)$. The parameter ϕ are the **variational parameters** and are used to characterize each member of the variational family. Our goal then is to find the value for ϕ that makes $q(z \mid \phi)$ as "close" to $p(z \mid x)$ as possible.

Variational inference uses the KL-divergence from $p(z \mid x)$ to $q(z \mid \phi)$ as a measure of "closeness":

$$KL(q(z \mid \phi) \parallel p(z \mid \phi)) := E_{Z \sim q}\left[\log \frac{q(Z \mid \phi)}{p(Z \mid x)}\right] \tag{1}$$

Thus, variational inference attempts to find

$$\hat{\phi} := \text{argmin}_{\phi}\, KL(q(z \mid \phi) \parallel p(z \mid x))$$

and then returns

$$q(z \mid \hat{\phi})$$

as the approximation to the posterior.

In general, there are many algorithms for performing variational inference, each with strengths and weaknesses depending on $p(z, x)$ and the surrogate family. Variational inference is a high-level strategy rather an explicit algorithm.

Details

At the core of variational inference is the use of Jensen's inequality:

$$E[f(X)] \leq f\left(E[X]\right)$$

where f is a convex function. We derive the variational inference algorithm by applying Jensen's Inequality on the log-marginal likelihood (also called the **evidence**):

$$\begin{aligned}
\log p(x) &= \log \int p(x, z)\, dz \\
&= \log \int p(x, z) \frac{q(z \mid \phi)}{q(z \mid \phi)}\, dz \\
&= \log \left(E_{Z \sim q} \left[\frac{p(x, Z)}{q(Z \mid \phi)} \right] \right) \\
&\geq E_{Z \sim q} \left(\log \frac{p(x, Z)}{q(Z \mid \phi)} \right) && \text{Jensen's Inequality} \\
&= E_{Z \sim q} \left[\log p(x, Z) \right] - E_{Z \sim q} \left[\log q(Z \mid \phi) \right]
\end{aligned}$$

This final quantity is called the **evidence lower bound** (ELBO) since it is provides a lower bound for the evidence. Variational inference finds ϕ that maximizes the ELBO. We show below that this also minimizes the KL-divergence in Equation 1, which is our true goal:

$$\begin{aligned}
KL(q(z \mid \phi) \,||\, p(z \mid x)) &= E_{Z \sim q} \left[\log \frac{q(Z \mid \phi)}{p(Z \mid x)} \right] \\
&= E_{Z \sim q} \left[\log q(Z \mid \phi) \right] - E_{Z \sim q} \left[\log p(Z \mid x) \right] \\
&= E_{Z \sim q} \left[\log q(Z \mid \phi) \right] - E_{Z \sim q} \left[\log p(Z \mid x) \right] \\
&= E_{Z \sim q} \left[\log q(Z \mid \phi) \right] - E_{Z \sim q} \left[\log \frac{p(Z, x)}{p(x)} \right] \\
&= E_{Z \sim q} \left[\log q(Z \mid \phi) \right] - E_{Z \sim q} \left[\log p(Z, x) \right] + E_{Z \sim q} \left[\log p(x) \right] \\
&= \log p(x) - \left(E_{Z \sim q} \left[\log p(x, Z) \right] - E_{Z \sim q} \left[\log q(Z \mid \phi) \right] \right) \\
&= \log p(x) - \text{ELBO}
\end{aligned}$$

In the last line, we see that the KL-divergence is the evidence minus the ELBO. Thus, in order to minimize the KL-divergence, we need to maximize the ELBO. (As another interesting note, we see that the difference between the evidence and the ELBO as given by Jensen's inequality is exactly the KL-divergence from $q(z \mid \phi)$ to $p(z \mid x)$).