Variational Inference and Mean Field

Variational Bayesian methods are a set of techniques to approximate posterior distributions in Bayesian Inference. First, we're trying to perform "Bayesian inference", which basically means given a model, we're trying to find distributions for the unobserved variables (either parameters or latent variables since they're treated the same). This problem usually involves hard-to-solve integrals with no analytical solution.

There are two main avenues to solve this problem. The first is to just get a point-estimate for each of the unobserved variables (either MAP or MLE) but this is not ideal since we can't quantify the uncertainty of the unknown variables (and is against the spirit of Bayesian analysis). The other aims to find a (joint) distribution of each unknown variable. With a proper distribution for each variable, we can do a whole bunch of nice Bayesian analysis like the mean, variance, 95% credible interval etc.

One good but relatively slow method for finding a distribution is to use MCMC to iteratively draw samples that eventually give you the shape of the joint distribution of the unknown variables. Another method is to use variational Bayes, which helps to find an approximation of the distribution in question. With variational Bayes, you only get approximation but it's in analytical form (read: easy to compute). So long as your approximation is pretty good, you can do all the nice Bayesian analysis you like, and the best part is it's relatively easy to compute.

Now that we know our problem, next thing we need to is define what it means to be a good approximation. In many of these cases, the Kullback-Leibler divergence (KL divergence) is a good choice, which is non-symmetric measure of the difference between two probability distributions P and Q. We'll discuss this in detail in the box below, but the setup will be Pas the true posterior distribution, and Qbeing the approximate distribution, and with a bit of math, we want to find an iterative algorithm to compute Q.
In the mean-field approximation (a common type of variational Bayes), we assume that the unknown variables can be partitioned so that each partition is independent of the others. Using KL divergence, we can derive mutually dependent equations (one for each partition) that define the shape of Q. The resultant Q function then usually takes on the form of well-known distributions that we can easily analyze. The leads to an easy-to-compute iterative algorithm (similar to the EM algorithm where we use all other previously calculated partitions to derive the current one in an iterative fashion.

To summarize, variational Bayes has these ideas:
+ The Bayesian inference problem of finding a posterior on the unknown variables (parameters and latent variables) is hard and usually can't be solved analytically.
+ Variational Bayes solves this problem by finding a distribution Q that approximates the true posterior P.
+ It uses KL-divergence as a measure of how well our approximation fits the true posterior.
+ The mean-field approximation partitions the unknown variables and assumes each partition is independent (a simplifying assumption).
+ With some (long) derivations, we can find an algorithm that iteratively computes the Q distributions for a given partition by using the previous values of all the other partitions.

Kullback-Leibler Divergence

Is a non-symmetric measure of the difference between two probability distributions $P$and Q. It is defined for discrete and continuous probability distributions as such:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

where p and q denote the densities of P and Q.

Entropy is the average amount of information or "surprise" for a probability distribution. Entropy is defined as for both discrete and continuous distributions:

$$H(P) = E_P[I_P(X)] = -\sum_{i=1}^{n} P(i)\log(P(i)) \tag{2}$$

$$H(P) = E_P[I_P(X)] = -\int_{-\infty}^{\infty} p(x)\log(p(x))dx$$

Thus, KL divergence can be viewed as this average extra-message length we need when we wrongly assume the probability distribution, using Q instead of P

$$D_{KL}(P||Q) = H(P,Q) - H(P)$$

$$D_{KL}(P||Q) = -\sum_{i=1}^{n} P(i)\log(Q(i)) + \sum_{i=1}^{n} P(i)\log(P(i))$$

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(i)\log\frac{P(i)}{Q(i)} \tag{3}$$

You can probably already see how this is a useful objective to try to minimize. If we have some theoretic minimal distribution P, we want to try to find an approximation Q that tries to get as close as possible by minimizing the KL divergence.

One thing to note about KL divergence is that it's not symmetric, that is, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The former is called forward KL divergence, while the latter is called reverse KL divergence. Let's start by looking at forward KL. Taking a closer look at equation 5, we can see that when P is large and $Q \to 0$, the logarithm blows up. This implies when choosing our approximate distribution Q to minimize forward KL divergence, we want to "cover" all the non-zero parts of Pas best we can.

Now, let's take a look at reverse KL, where P is still our theoretic distribution we're trying to match and Q is our approximation:

$$D_{KL}(Q||P) = \sum_{i=1}^{n} Q(i)\log\frac{Q(i)}{P(i)} \tag{4}$$

From Equation 6, we can see that the opposite situation occurs. If P is small, we want Q to be (proportionally) small too or the ratio might blow up. Additionally, when P is large, it doesn't cause us any particular problems because it just means the ratio is close to 0.

In our use of KL divergence, we'll be using reverse KL divergence, not only because of the nice properties above, but for the more practical reason that the math works out nicely

From KL divergence to Optimization

Remember what we're trying to accomplish: we have some intractable Bayesian inference problem $P(\theta|X)$ we're trying to compute, where θ are the unobserved variables (parameters or latent variables) and $X$are our observed data. We could try to compute it directly using Bayes theorem (continuous version, where p is the density of distribution P):

$$p(\theta|X) = \frac{p(X,\theta)}{p(X)}$$

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_{-\infty}^{\infty} p(X|\theta)p(\theta)d\theta} \tag{5}$$

$$p(\theta|X) = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal likelihood}}$$

However, this is generally difficult to compute because of the marginal likelihood (sometimes called the evidence). But what if we didn't have to directly compute the marginal likelihood and instead only needed the likelihood.

This idea leads us to the two commonly used methods to solve Bayesian inference problems: MCMC and variational inference. MCMC in general it's quite slow since it involves repeated sampling but your approximation can get arbitrarily close to the actual distribution (given enough time). Variational inference on the other hand is a strict approximation that is much faster because it is an optimizing problem. It also can quantify the lower bound on the marginal likelihood, which can help with model selection.

Now going back to our problem, we want to find an approximate distribution Q that minimizes the (reverse) KL divergence. Starting from reverse KL divergence, let's do some manipulation to get to an equation that's easy to interpret (using continuous version here), where our approximate density is $q(\theta)$and our theoretic one is $p(\theta|X)$:

$$\begin{aligned} D_{KL}(Q||P) &= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta|X)} d\theta \\ D_{KL}(Q||P) &= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta, X)} d\theta + \int_{-\infty}^{\infty} q(\theta) \log p(X) d\theta \qquad (6) \\ D_{KL}(Q||P) &= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta, X)} d\theta + \log p(X) \end{aligned}$$

Where we're using Bayes theorem on the second line and the RHS integral simplifies because it's simply integrating over the support of $q(\theta)$ ($\log p(X)$ is not a function of θ so it factors out). Rearranging we get:

$$\log p(X) = D_{KL}(Q||P) - \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta, X)} d\theta$$

$$\log p(X) = D_{KL}(Q||P) + \mathcal{L}(Q) \qquad (7)$$

where $\mathcal{L}$ is called the (negative) **variational free energy**, NOT the likelihood. Recall that the evidence on the LHS is constant (for a given model), thus if we maximize the variational free energy $\mathcal{L}$, we minimize (reverse) KL divergence as required.

This is the crux of variational inference: we don't need to explicitly compute the posterior (or the marginal likelihood), we can solve an optimization problem by finding the right distribution Q that best fits our variational free energy. Notice that we don't need to compute the marginal likelihood either, this is a big win because the likelihood and prior are usually easily specified with the marginal likelihood intractable. Note that we need to find a function, not just a point, that maximizes $\mathcal{L}$, which means we need to use variational calculus, hence the name **"Variational Bayes"**.

The Mean-Field Approximation

Before we try to derive the functional form of our Q functions, let's just explicitly state some of our notation because it's going to get a bit confusing. In the previous section, I used θ to represent the unknown variables. In general, we can have N unknown variables so $\theta = (\theta_1, \dots, \theta_N)$ and Equation 8 and 9 will have multiple integrals (or summations for discrete variables), one for each θ_i. I'll use θ to represent $\theta_1, \dots, \theta_N$ where it is clear just to reduce the verbosity and explicitly write it out when we want to do something special with it.

Okay, so now that's cleared up, let's move on to the mean-field approximation. The approximation is a simplifying assumption for our Q distribution, which partitions the variables into independent parts (I'm just going to show one variable per partition but you can have as many per partition as you want):

$$p(\theta|X) \approx q(\theta) = q(\theta_1, \dots, \theta_n) = \prod_{i=1}^{N} q_i(\theta_i) \tag{8}$$

From Equation 8, we can plug it back into our variational free energy $\mathcal{L}$and try to derive the functional form of $q_j$using variational calculus. Let's start with $\mathcal{L}$and try to re-write it isolating the terms for $q_j(\theta_j)$in hopes of taking a functional derivative afterwards to find the optimal form of the function. Note that $\mathcal{L}$is a functional that depends on our approximate densities $q_1, \dots, q_N$.

$$\begin{aligned}
\mathcal{L}[q_1, \dots, q_N] &= -\int_{\theta_1,\dots,\theta_N} [\prod_{i=1}^{N} q_i(\theta_i)] \log \frac{[\prod_{k=1}^{N} q_k(\theta_k)]}{p(\theta, X)} d\theta_1 \dots d\theta_n \\
\mathcal{L}[q_1, \dots, q_N] &= \int_{\theta_1,\dots,\theta_N} [\prod_{i=1}^{N} q_i(\theta_i)] \big[\log p(\theta, X) - \sum_{k=1}^{N} \log q_k(\theta_k)\big] d\theta_1 \dots d\theta_n \\
\mathcal{L}[q_1, \dots, q_N] &= \int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \big[\log p(\theta, X) - \sum_{k=1}^{N} \log q_k(\theta_k)\big] d\theta_1 \dots d\theta_n \qquad (9)
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}[q_1, \dots, q_N] = &\int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \log p(\theta, X) d\theta_1 \dots d\theta_n \\
&- \int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \sum_{k=1}^{N} \log q_k(\theta_k) d\theta_1 \dots d\theta_n
\end{aligned}$$

So far, we've just factored out $q_j(\theta_j)$ and multiplied out the inner term $\log p(\theta, X) - \sum_{i=k}^{N} \log q_k(\theta_k)$. In anticipation of the next part, we'll define some notation for an expectation across all variables except j as:

$$E_{m|m\neq j}[\log p(\theta, X)] = \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \log p(\theta, X) d\theta_1 \dots, d\theta_{j-1}, d\theta_{j+1}, \dots, d\theta_n \tag{10}$$

which you can see is just an expectation across all variables except for j. Continuing on from Equation 9 using this expectation notation and expanding the second term out:

$$\begin{aligned}
\mathcal{L}[q_1, \dots, q_N] = &\int_{\theta_j} q_j(\theta_j) E_{m|m\neq j}[\log p(\theta, X)] d\theta_j - \int_{\theta_j} q_j(\theta_j) \log q_j(\theta_j) \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] d\theta_1 \dots d\theta_n \\
&- \int_{\theta_j} q_j(\theta_j) d\theta_j \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \sum_{k\neq j} \log q_k(\theta_k) d\theta_1 \dots, d\theta_{j-1}, d\theta_{j+1}, \dots, d\theta_n
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}[q_1, \dots, q_N] = &\int_{\theta_j} q_j(\theta_j) E_{m|m\neq j}[\log p(\theta, X)] d\theta_j - \int_{\theta_j} q_j(\theta_j) \log q_j(\theta_j) d\theta_j \\
&- \int_{\theta_{m|m\neq j}} [\prod_{i\neq j} q_i(\theta_i)] \sum_{k\neq j} \log q_k(\theta_k) d\theta_1 \dots, d\theta_{j-1}, d\theta_{j+1}, \dots, d\theta_n \qquad (11)
\end{aligned}$$

$$\mathcal{L}[q_1, \dots, q_N] = \int_{\theta_j} q_j(\theta_j) \big[E_{m|m\neq j}[\log p(\theta, X)] - \log q_j(\theta_j)\big] d\theta_j \quad - G[q_1, \dots, q_{j-1}, q_{j+1}, \dots, q_N]$$

where we're integrating probability density functions over their entire support in a couple of places, which simplifies a few of the expressions to 1. At the end, we have a functional that consists of a term made up only of $q_j(\theta_j)$ and $E_{m|m\neq j}[\log p(\theta, X)]$, and another term with all the other q_i functions.

Putting together the Lagrangian for Equation 11, we get:

$$\mathcal{L}[q_1, \ldots, q_N] - \sum_{i=1}^{N} \lambda_i \int_{\theta_i} q_i(\theta_i) d\theta_i \tag{12}$$

where the terms in the summation are our usual probabilistic constraints that the $q_i(\theta_i)$ functions must be probability density functions. Taking the functional derivative of Equation 12 with respect to $q_j(\theta_j)$ we get:

$$\begin{aligned} \frac{\delta \mathcal{L}[q_1, \ldots, q_N]}{\delta q_j(\theta)} &= \frac{\partial}{\partial q_j}\left[q_j(\theta_j)\left[E_{m|m\neq j}[\log p(\theta, X)] - \log q_j(\theta_j)\right] - \lambda_j q_j(\theta_j)\right] \\ &= E_{m|m\neq j}[\log p(\theta, X)] - \log q_j(\theta_j) - 1 - \lambda_j \end{aligned} \tag{13}$$

In this case, the functional derivative is just the partial derivative with respect to $q_j(\theta_j)$ of what's "inside" the integral. Setting to 0 and solving for the form of $q_j(\theta_j)$:

$$\begin{aligned} \log q_j(\theta_j) = E_{m|m\neq j}[\log p(\theta, X)] - 1 - \lambda_j = E_{m|m\neq j}[\log p(\theta, X)] + \text{const} \\ q_j(\theta_j) = \frac{e^{E_{m|m\neq j}[\log p(\theta, X)]}}{Z_j} \end{aligned} \tag{14}$$

where Z_j is a normalization constant. The constant isn't too important because we know that $q_j(\theta_j)$ is a density so usually we can figure it out after the fact.

Equation 14 finally gives us the functional form (actually a template of the functional form). What usually ends up happening is that after plugging in $E_{m|m\neq j}[\log p(\theta, X)]$, the form of Equation 16 matches a familiar distribution (e.g. Normal, Gamma etc.), and the normalization constant Z can be derived by inspection. We'll see this play out in the next section.

Taking a step back, let's see how this helps us accomplish our goal. Recall, we wanted to maximize our variational free energy $\mathcal{L}$ (Equation 7), which in turn finds a $q(\theta)$ that minimizes KL divergence to the true posterior $p(\theta|X)$. Using the mean-field approximation, we broke up $q(\theta)$ (Equation 8) into partitions $q_j(\theta_j)$, each of which is defined by Equation 14.

However, the $q_j(\theta_j)$'s are interdependent when minimizing them. That is, to compute the optimal $q_j(\theta_j)$, we need to know the values of all the other $q_i(\theta_i)$ functions (because of the expectation $E_{m|m\neq j}[\log p(\theta, X)]$). This suggests an iterative optimization algorithm:

1. Start with some random values for each of the parameters of the $q_j(\theta_j)$ functions.
2. For each q_j, use Equation 14 to minimize the overall KL divergence by updating $q_j(\theta_j)$ (holding all the others constant).
3. Repeat until convergence.

Notice that in each iteration, we are lowering the KL divergence between our Q and P distributions, so we're guaranteed to be improving each time. Of course in general we won't converge to a global maximum but it's a heck of a lot easier to compute than MCMC.

This document has been created by Naman Deep Singh using the original post [1]*. All the derivations and text belongs to the author of this post.*

[1] http://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/