A Primer on Variational Inference

Harrison Edwards

February 25, 2016

1 Introduction

This informal document is designed to refresh your memory about variational inference. I like to use uppercase italics like X to denote random variables and lower case letters like x to denote a particular value that the random variable X has taken.

2 The Problem

We want to build a nice probabilistic model for our data so we gather our ingredients:

- Observed data $X \in \mathbb{R}^d$
- Latent variables or parameters $Z \in \mathbb{R}^k$
- A prior $p(Z)$
- A conditional model $p(X|Z)$

As a researcher, we will often begin our work with some ingenious scheme for $p(X|Z)$ using our intuition and/or domain knowledge (or just use a giant neural network), for instance we might decide to model documents X as being a mixture of latent topics Z.

It is often less easy to have an intuition about our prior $p(Z)$, but once we have specified $p(X|Z)$ we often gain some insight into how the prior affects the model, and can then choose $p(Z)$ accordingly. For instance in the topic model we can encode our beliefs about how many topics are present in the typical document.

The next thing we will want to do is ask, for a given x, what z generated it? In other words we want to know about $p(z|x)$. That's easy, you will think, I just apply Bayes' rule

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{1}$$

but where does $p(x)$ come from? Then you remember that you are a good little Bayesian, and that all good Bayesians integrate out their parameters

$$p(x) = \int_z p(x|z)p(z)\, dz \tag{2}$$

and now you are sad because your ingenious scheme for $p(X|Z)$ has meant that you have absolutely no idea how to do this integral.

3 A General Approach

As we have said, the problem is to compute $p(Z|X)$, but this may have an arbitrarily complex form even for simple/easy to evaluate $p(Z)$ and $p(Z|X)$. Instead of trying to compute $p(Z|X)$ exactly, one strategy is to pick a family of easy-to-work-with distributions $q(Z|\phi)$ and choose ϕ such that $q(Z|\phi)$ approximates $p(Z|X)$ in some sense. That is

$$q(z|x) = \min_{\phi} G(q(z|\phi), p(z|x)),$$

where G is some function that measures how dissimilar two distributions are. Choices of different families of distributions for q and choices of G lead to different methods. Variational inference chooses G to be the Kullback-Leibler(KL) divergence. Laplacian methods choose q to be a Gaussian centered at the mode of $p(Z|x)$ and G to be the distance between the Hessians of the pdfs.

4 The Variational Solution

4.1 Useful Notions from Information Theory

4.1.1 Entropy

Given a pdf/pmf p we define the entropy / differential entropy to be

$$H(p) = -\int_x p(x) \log p(x)\, dx. \tag{3}$$

The quantity $-\log p(x)$ is called the *surprisal*, and is meant to quantify how 'surprised' you should be if you sampled x using the probability distribution p. The entropy then is the expected surprise and is a measure of how uncertain or spread out a distribution is. When the logarithm is base 2 the unit of entropy is *bits*, and when we use the natural logarithm the unit is *nats*. Also note that by convention we define $0 \cdot \log 0$ to be zero.

For example: if p is discrete taking values $1, \ldots, k$, then the entropy is *maximised* when $p(1) = \cdots = p(k) = \frac{1}{k}$, since we have no reason to expect any particular outcome more than the others. On the other hand, the entropy would be *minimised* when $p(i) = 1$ for some i between 1 and k. This is because there is no surprise at all, we are certain of what will happen when we sample x according to p.

4.1.2 Cross-Entropy

The entropy is defined as the expectation under p of the surprisal $-\log p(x)$, but what happens if you change the distribution in the surprisal? The name of the answer is the *cross-entropy*, defined

$$H(p, q) = -\int_x p(x) \log q(x)\, dx. \tag{4}$$

The way I like to think about this is that you *believe* that the samples are coming from q, and hence when you see x you are $-\log q(x)$ surprised, but *in reality* the samples are coming from p and so the amount you are surprised on average is different from the true entropy $H(p)$. Intuitively, if p is different to q, then your expectations are incorrect and so you will be more surprised on average. This is both intuitively true and actually true, and so important it has a name: *Gibb's Inequality*

$$H(p, q) \geq H(p), \tag{5}$$

or equivalently

$$H(q, p) \geq H(q). \tag{6}$$

This can be shown using Jensen's Inequality.

4.1.3 KL-Divergence

When your beliefs don't match reality, you are, on average, more surprised by events than you should be, but how much more? The name for this difference is called the Kullback-Leibler(KL)-divergence, defined

$$D_{KL}\left(p \| q\right) = H(p, q) - H(p). \tag{7}$$

By Gibb's Inequality, the KL-divergence is always non-negative, and it can also be shown that $D_{KL}\left(q \| p\right) = 0$ if and only if $p = q$ almost everywhere.

One very important thing to note is that the KL-divergence is *not symmetric*, that is $D_{KL}\left(p \| q\right)$ does not in general equal $D_{KL}\left(q \| p\right)$. Both are valid ways of measuring dissimilarity between distributions, but they emphasise different aspects of the discrepancy. If you adjust your beliefs q to minimise $D_{KL}\left(p \| q\right)$, you are an *optimist*: you think that it is very important to believe everything that is true (even it means believing in some things are are false). Conversely, if you instead adjust your beliefs q to minimise $D_{KL}\left(q \| p\right)$, you are a *sceptic*: you think it is very important not to believe in anything false (even if it means not believing in some things that turn out to be true). So which are you?

5 Variational Inference

In life it is often observed that it is easier to be a sceptic than an optimist, and probability is no exception. In variational inference we choose $q(Z)$ to minimise $D_{KL}\left(q(Z|x) \| p(Z|x)\right)$ because

$$D_{KL}\left(p(Z|x) \| q(Z|x)\right) = -\int_z p(z|x) \log q(z|x)\, dz + \int_z p(z|x) \log p(z|x)\, dz$$

is *hard* to calculate, because we don't know how to sample from $p(Z|x)$. We know how to sample from $q(Z|x)$ - we chose q specifically because it is easy to work with. So instead we want to minimise

$$D_{KL}\left(q(Z|x) \| p(Z|x)\right) = -\int_z q(z|x) \log p(z|x)\, dz + \int_z q(z|x) \log q(z|x)\, dz$$

with respect to q. Let's play with this quantity and see where it gets us:

$$\begin{aligned}
D_{KL}\left(q(Z|x) \| p(Z|x)\right) &= -\int_z q(z|x) \log p(z|x)\, dz + \int_z q(z|x) \log q(z|x)\, dz \\
&= -\int_z q(z|x) \log p(z, x)\, dz + \int_z q(z|x) \log p(x)\, dz + \int_z q(z|x) \log q(z|x)\, dz \\
&= -\int_z q(z|x) \log p(z, x)\, dz + \log p(x) + \int_z q(z|x) \log q(z|x)\, dz
\end{aligned}$$

We have learned that

$$\log p(x) = D_{KL}\left(q(Z|x) \| p(Z|x)\right) + \int_z q(z|x) \log p(z,x)\, dz - \int_z q(z|x) \log q(z|x)\, dz \tag{8}$$

and observe that by using the fact that KL-divergences are non-negative we get that

$$\log p(x) \geq \int_z q(z|x) \log p(z,x)\, dz - \int_z q(z|x) \log q(z|x)\, dz. \tag{9}$$

This term is often called the *free energy*, because physics. Other people call it the *ELBO*, the *evidence based lower bound*. This is more sane. To see this note that

$$\begin{aligned}
\text{ELBO}\,(x) &= \int_z q(z|x) \log p(z,x)\, dz - \int_z q(z|x) \log q(z|x)\, dz \\
&= \int_z q(z|x) \log p(x|z)\, dz + \int_z q(z|x) \log p(z)\, dz - \int_z q(z|x) \log q(z|x)\, dz \\
&= \int_z q(z|x) \log p(x|z)\, dz - D_{KL}\left(q(Z|x) \| p(Z)\right)
\end{aligned}$$

so we have the KL-divergence between our approximate $q(Z|X)$ and our prior $p(Z)$, and the expected log probability of the data given z, where the expectation is taken with respect to $q(Z|x)$. To maximise the ELBO we must make the z's $q(Z|x)$ generates explain the data, but we also have to choose $q(Z|x)$ so that it is close to the prior, which keeps us honest.

Minimising $D_{KL}\left(q(Z|x) \| p(Z|x)\right)$ is then equivalent to maximising a lower bound on $\log p(x)$ which means maximising the ELBO. But is maximising the ELBO any easier to do? Yes! We know how to calculate $p(x|z)$, that was our initial ingenious scheme, we also know how to sample from $q(Z|x)$, therefore we can estimate $\int_z q(z|x) \log p(x|z)\, dz$. We can also calculate $p(z)$, therefore we can estimate $D_{KL}\left(q(Z|x) \| p(Z)\right)$, and if we were particularly cunning we might have chosen q so that we can evaluate $D_{KL}\left(q(Z|x) \| p(Z)\right)$ analytically, for instance if $q(Z)$ and $p(Z)$ are both Gaussians.

References