Tutorial: Stochastic Variational Inference

David Madras

University of Toronto

March 16, 2017

Variational Inference (VI) - Setup

- Suppose we have some data x, and some latent variables z (e.g. Mixture of Gaussians)
- We're interested in doing posterior inference over z
- This would consist of calculating:

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

- The numerator is easy to compute for given z, x
- The denominator is, in general, intractable

The Variational Distribution

- Rather than calculate the posterior $p(z|x)$ exactly, we can approximate it with some distribution q
- The approximating distribution q is called the *variational distribution*
- We'll choose q to have some nice form, so that we can feasibly calculate $q(z|x)$

KL Divergence

- Since q is intended to approximate p, we want them to be as similar as possible
- We can measure this in many ways - the most common is KL divergence:

$$\begin{aligned} KL(q(z|x)||p(z|x)) &= \int_z q(z|x) \log \frac{q(z|x)}{p(z|x)} \\ &= \mathbb{E}_q \log \frac{q(z|x)}{p(z|x)} \end{aligned} \tag{2}$$

- Note, however, we can't calculate this expression, since we don't have $p(z|x)$ - that's the whole point of what we're doing!

The ELBO

- However, we can re-write this KL divergence as follows:

$$
\begin{aligned}
&KL(q(z|x)||p(z|x)) \\
&\quad = \mathbb{E}_q \log \frac{q(z|x)}{p(z|x)} \\
&\quad = \mathbb{E}_q \log q(z|x) - \mathbb{E}_q \log p(z|x) \\
&\quad = \mathbb{E}_q \log q(z|x) - \mathbb{E}_q[\log p(z,x) - \log p(x)] \\
&\quad = \mathbb{E}_q[\log q(z|x) - \log p(z,x)] + \log p(x) \\
&\quad = -\mathbb{E}_q[\log p(z,x) - \log q(z|x)] + \log p(x)
\end{aligned} \tag{3}
$$

- The second term here is a constant, $\log p(x)$
- The first term is called the "ELBO" - Evidence Lower BOund
- So maximizing the ELBO is equivalent to minimizing the KL divergence between the posteriors!

VI notes

- The ELBO is important in generative modelling, a field of machine learning which is very popular at the moment
- It is how variational autoencoders (VAEs) work - the ELBO is a lower bound on $\log p(x)$
- It is often written with two terms: the "reconstruction loss" and a KL divergence with a prior over the latent variables:

$$ELBO = \mathbb{E}_q[\log p(x|z)] - KL(q(z|x)||p(z)) \tag{4}$$

- Another nice explanation of this material from David Blei (just the first few pages):
  - https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf

Reparametrization Trick

- In practice, we have to do some things to make VI work
- One important thing is the reparametrization trick
- We can learn $q(z|x)$ in a very flexible way by using a deep neural network to output the parameters of q
- Our network takes x as input, and outputs the parameters e.g. a mean and standard deviation if q is a Gaussian
- This is great because we can now combine all the power of deep learning with the structure of graphical models
- In the Gaussian case, we can output $\log \sigma$, which is more stable, and ensures $\sigma > 0$
- Remember we estimate the ELBO by sampling from q

Other Tricks

- Can slowly anneal in the KL term over the first n epochs
  - Put a coefficient on the KL term, keep it small at the beginning of training, and slowly increase it to 1 (as in https://arxiv.org/abs/1511.06349)
- Learning a global parameter's variance can have very high variance
- Rather than letting $\sigma = \exp(x)$ as mentioned above, we can use a softplus instead: $\sigma = \log(1 + \exp(x))$ (as in https://arxiv.org/abs/1505.05424)

Thanks!

- Any questions?