Expectation Maximization and Variational Inference

Xiaobao Wu
Tsinghua University
xiaobao.wu1996@gmail.com

1 Expectation Maximization

The advantages of EM are

1. no need to tune parameters
2. easy to program
3. more elegant

We assume $p(z|x;\theta^{old})$ is tractable.

$$Q(\theta, \theta^{old}) = \mathbb{E}_{p(z|x,\theta^{old})}\left[\log p(x,z)\right] = \sum_z p(z|x,\theta^{old}) logp(x,z;\theta) \tag{1}$$

which can be seen as the expectation of the complete-data log likelihood.

E-step: compute $p(z|x;\theta^{old})$ given θ^{old}.

M-step: $\theta^{new} = \mathrm{argmax}_\theta Q(\theta, \theta^{old})$.

2 General EM

Also known as Varitional EM. We use $q(z)$ instead of $q(z|x,\theta)$.

E-step with θ fixed:

$$q(z) = argmax_q \mathbb{E}_{q(z)}[\log p(x,z|\theta)] \tag{2}$$

M-step with $q(z)$ fixed:

$$\theta^{new} = argmax_\theta \mathbb{E}_{q(z)}[\log p(x,z|\theta)] \tag{3}$$

3 Varitional Inference

There is a difference between $K(q\|p)$ and $K(p\|q)$.

$$K(q(z)|\,p(z)) = \int q(z) \log \frac{p(z)}{q(z)} dz \tag{4}$$

This KL divergence is large in the region where $p(z)$ is zero unless $q(z)$ is also zero. Thus, $q(z)$ tends to become small where $p(z)$ is small. Conversely, $q(z)$ is nonzero where $p(z)$ is nonzero.

If $p(z|x;\theta^{old})$ is intractable, we approximate the posterior probability using a simpler model, which comes to Varitional Inference methods.

The ELBO is

$$\begin{aligned}\mathcal{L}(q(z)) &= \mathbb{E}_{q(z)}\left[\log p(x,z) - \log q(z)\right] && (5)\\ &= \mathbb{E}_{q(z)}\left[\log p(x,z)\right] - \int q(z) \log q(z) dz && (6)\end{aligned}$$

Thus, traditional EM algorith is identical with Varitional Inference when $q(z) = q(z|x, \theta)$.

4 Stochastic Gradient Varitional Inference

We take $argmax\mathcal{L}$ as an optimization problem. Here, $q_\phi(z)$ is same as $q_\phi(z|x)$.

$$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi \mathbb{E}_{q_\phi(z)} \left[\log p_\theta(x, z) - \log q_\phi(z)\right] \tag{7}$$

$$= \mathbb{E}_{q_\phi(z)} \left[(\log p_\theta(x, z) - \log q_\phi(z)) \nabla_\phi \log q_\phi(z)\right] \tag{8}$$

We now can use MC to approximate it:

$$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{i=1}^{L} (\log p_\theta(x, z^{(l)}) - \log q_\phi(z^{(l)})) \nabla_\phi \log q_\phi(z^{(l)}) \tag{9}$$

Due to the property of log function within $(0, 1]$, the variance of $(\log p_\theta(x, z) - \log q_\phi(z)) \nabla_\phi \log q_\phi(z)$ will be very large.

We can use reparametrization trick to alleviate it. There are also other methods like REINFORCE algorithm. We assume $z = g_\phi(\epsilon, x), \epsilon \sim p(\epsilon)$ and we have $q(z|x)dz = p(\epsilon)d\epsilon$

$$\nabla_\phi \mathcal{L}(\phi) = \mathbb{E}_{p(\epsilon)} \left[\nabla_\phi (\log p_\theta(x, z) - \log q_\phi(z)))\right] \tag{10}$$

$$= \mathbb{E}_{p(\epsilon)} \left[(\nabla_z (\log p_\theta(x, z) - \log q_\phi(z)))\, \nabla_\phi g_\phi(\epsilon, x)\right] \tag{11}$$

$$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{i=1}^{L} \nabla_z (\log p_\theta(x, z) - \log q_\phi(z))) \nabla_\phi g_\phi(\epsilon^{(l)}, x) \tag{12}$$

where $z = g_\phi(\epsilon^{(l)}, x)$.

Update as

$$\phi^{(t+1)} = \phi^{(t)} + \lambda \nabla_\phi (\mathcal{L}(\phi)) \tag{13}$$