Expectation-maximization

He Li, NYU Stern

Last Update: November 14, 2018

1 Latent Variable Model

Let X denote the observable variable, and let Z denote the latent variable. The probability model is $p(x, z|\theta)$. If Z is instead observable, then MLE takes the maxima of *the complete likelihood*:

$$l_c(\theta; x, z) \triangleq \log p(x, z|\theta) \tag{1.1}$$

However, here MLE should takes the maxima of *the incomplete likelihood*

$$l(\theta; x) \triangleq \log p(x|\theta) = \log \int_Z p(x, z|\theta)\,\mathrm{d}z \tag{1.2}$$

where we consider only X's marginal distribution. We can also use an candidate distribution to remove the randomness over complete likelihood:

$$\langle l_c(\theta; x, z)\rangle_q \triangleq \int_Z q(z|x, \theta) \log p(x, z|\theta)\,\mathrm{d}z \tag{1.3}$$

where $\langle l_c(\theta; x, z)\rangle_q$ is called *expected complete likelihood.* Now

$$\begin{aligned} l(\theta; x) &= \log p(x|\theta) \\ &= \log \int_Z p(x, z|\theta)\,\mathrm{d}z \\ &= \log \int_Z q(z|x, \theta) \frac{p(x, z|\theta)}{q(z|x, \theta)}\,\mathrm{d}z \\ &\geq \int_Z q(z|x, \theta) \log \frac{p(x, z|\theta)}{q(z|x, \theta)}\,\mathrm{d}z \\ &\triangleq \mathcal{L}(q, \theta) \end{aligned} \tag{1.4}$$

2 Coordinate Ascent Aspect

The Expectation-maximization algorithm can be seen as a coordinate ascent algorithm on the function $\mathcal{L}(q, \theta)$:

$$\textbf{(E step)} \qquad q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)}) \tag{2.1}$$

$$\textbf{(M step)} \qquad \theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta) \tag{2.2}$$

Notice that in E step, $q^{(t+1)}(z|x) = p(z|x, \theta^{(t)})$. One can easily verify that since $\mathcal{L}(p(z|x, \theta^{(t)}), \theta^{(t)}) = l(\theta^{(t)}, x)$, which is its upper bound.

Therefore, in E step, we actually close the gap between $\mathcal{L}(q, \theta^{(t)}))$ and incomplete likelihood $l(\theta^{(t)}), x)$. In M step, we increase the lower bound of incomplete likelihood, and since we close the gap between incomplete likelihood and the lower bound, we also increase the incomplete likelihood at the same time.

Therefore, another way to write the EM algorithm is:

$$\textbf{(E step)} \qquad q^{(t+1)} = p(z|x, \theta^{(t)}) \tag{2.3}$$

$$\textbf{(M step)} \qquad \theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(q^{(t+1)}, \theta) \tag{2.4}$$

Note that in M step, from the form in (1.4), we are actually maximize the expected complete likelihood $\langle l_c(\theta; x, z)\rangle_q = \int_Z q(z|x, \theta) \log p(x, z|\theta) \,\mathrm{d}z$.

We also notice without proof that this gap is actually a KL divergence.

$$l(\theta, x) - \mathcal{L}(q, \theta) = \mathrm{D}(q(z|x) \| p(z|x, \theta)) \tag{2.5}$$

3 Alternating Minimization

We first introduce the empirical distribution:

$$\tilde{p}(x) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta(x, x_n) \tag{3.1}$$

where $x,\tilde{x}_n$ is a Kronecker delta function. It can be shown that

$$\begin{aligned}
\int_X \tilde{p}(x) \log p(x|\theta) &= \int_X \frac{1}{N} \sum_{i=1}^{N} \delta(x, x_n) \log p(x|\theta) \,\mathrm{d}x \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_X \delta(x, x_n) \log p(x|\theta) \,\mathrm{d}x \\
&= \frac{1}{N} \sum_{i=1}^{N} \log p(x_n|\theta) \,\mathrm{d}x \\
&= \frac{1}{N} l(\theta|D)
\end{aligned} \tag{3.2}$$

which is the log likelihood of data. Further,

$$D(\tilde{p}(x) \| p(x|\theta)) = \int_X \tilde{p}(x) \log \tilde{p}(x) \,\mathrm{d}x - \frac{1}{N} l(\theta|D) \tag{3.3}$$

Since the first term is independent of θ, therefore, the maximizing value of θ with respect to the log likelihood is equal to the minimizing value of θ with respect to the KL divergence between empirical distribution and $p(x|\theta)$.

Now head back to the EM algorithm, it can be shown that

$$D(\tilde{p}(x) \| p(x|\theta)) \leq D\left(\tilde{p}(x) q(z|x) \| p(x, z|\theta)\right) \tag{3.4}$$

since $\int_X \tilde{p}(x) \log \tilde{p}(x) \,\mathrm{d}x$ is independent of q and θ, therefore it is equivalent to minimize $D(\tilde{p}(x) \| p(x|\theta))$. Therefore, the alternative EM algorithm is:

$$\textbf{(E step)} \qquad q^{(t+1)} = \arg\max_{q} D\left(\tilde{p}(x) q(z|x) \| p(x, z|\theta^{(t)})\right) \tag{3.5}$$

$$\textbf{(M step)} \qquad \theta^{(t+1)} = \arg\max_{\theta} D\left(\tilde{p}(x) q^{(t+1)}(z|x) \| p(x, z|\theta)\right) \tag{3.6}$$