Expectation Maximization 01 Algorithm Convergence

Chen Gong

17 December 2019

Expectation Maximization (EM) 算法，是用来解决具有隐变量的模型的概率计算问题。在比较简单的情况中，我们可以直接得出我们想要求得的参数的解析解，比如：MLE: $p(X|\theta)$。我们想要求解的结果就是：

$$\theta_{MLE} = \arg\max_{\theta} \log p(X|\theta) \tag{1}$$

然而一旦问题变得复杂起来以后，就不是这么简单了，特别是引入了隐变量之后。

1 EM 算法简述

实际上，EM 算法的描述也并不是很难，我们知道，通常我们想求的似然函数为 $p(X|\theta)$。引入隐变量之后，原式就变成了：

$$p(X|\theta) = \int \log p(X, Z|\theta) p(Z|X, \theta^{(t)}) dZ \tag{2}$$

EM 算法是一种迭代的算法，我们的目标是求：

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_{\theta} \int_Z \log p(X, Z|\theta) p(Z|X, \theta^{(t)}) dZ \\ &= \arg\max_{\theta} \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}[\log p(X, Z|\theta)] \end{aligned} \tag{3}$$

也就是找到一个更新的参数 θ，使得 $\log p(X, Z|\theta)$ 出现的概率更大。

2 EM 算法的收敛性

我们想要证的是当 $\theta^{(t)} \longrightarrow \theta^{(t+1)}$ 时，有 $\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$。这样才能说明我们的每次迭代都是有效的。

$$\log p(X|\theta) = \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} = \log p(X, Z|\theta) - \log p(Z|X, \theta) \tag{4}$$

下一步，则是同时对两边求关于 $p(Z|X, \theta^{(t)})$ 的期望。

左边：

$$\begin{aligned} \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}[\log p(X|\theta)] &= \int_Z p(Z|X, \theta^{(t)} \log p(X|\theta) dZ \\ &= \log p(X|\theta) \int_Z p(Z|X, \theta^{(t)}) dZ \\ &= \log p(X|\theta) \cdot 1 = \log p(X|\theta) \end{aligned} \tag{5}$$

右边：

$$\underbrace{\int_Z p(Z|X,\theta^{(t)})\log p(X,Z|\theta)dZ}_{Q(\theta,\theta^{(t)})} - \underbrace{\int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta)dZ}_{H(\theta,\theta^{(t)})} \tag{6}$$

大家很容易就观察到，$Q(\theta,\theta^{(t)})$ 就是我们要求的 $\theta^{(t+1)} = \arg\max_\theta \int_Z p(X,Z|\theta)p(Z|X,\theta^{(t)})dZ$。那么，根据定义，我们可以很显然的得到：$Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta,\theta^{(t)})$。当 $\theta = \theta^{(t)}$ 时，等式也是显然成立的，那么我们可以得到：

$$Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)}) \tag{7}$$

这时，大家想一想，我们已经得到了 $Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)})$ 了。如果，$H(\theta^{(t+1)},\theta^{(t)}) \leq H(\theta^{(t)},\theta^{(t)})$。我们就可以很显然的得出，$\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$ 了。

证明：

$$\begin{aligned} H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) &= \int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta^{(t+1)})dZ - \int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta^{(t)})dZ \\ &= \int_Z p(Z|X,\theta^{(t)})\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}dZ \\ &= -KL(p(Z|X,\theta^{(t)})||p(Z|X,\theta^{(t+1)})) \leq 0 \end{aligned} \tag{8}$$

或者，我们也可以使用 Jensen inequality。很显然，log 函数是一个 concave 函数，那么有 $\mathbb{E}[\log X] \leq \log[\mathbb{E}[X]]$，那么：

$$\begin{aligned} \int_Z p(Z|X,\theta^{(t)})\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}dZ =& \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}\left[\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right] \\ \leq& \log\left[\mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}\left[\frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right]\right] \\ =& \log\left[\int_Z p(Z|X,\theta^{(t)})\left[\frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right]dZ\right] \\ =& \log\int_Z p(Z|X,\theta^{(t+1)})dZ \\ =& 0 \end{aligned} \tag{9}$$

所以，从两个方面我们都证明了，$\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$。那么，经过每次的迭代，似然函数在不断的增大。这就证明了我们的更新是有效的，也证明了算法是收敛的。