EM 算法

Li Liang*

1 EM 算法

考虑有观测样本数据 X，需要找出样本的模型参数 θ，似然函数可以写成：

$$\ell(\theta) = \log p(X \mid \theta) \tag{1}$$

利用最大化似然法可以估计 $\theta = \arg\max_{\theta} \ell(\theta)$。但在某些任务中，观测样本的概率分布有些复杂，很难直接采用 $p(X \mid \theta)$ 计算，需要引入隐变量，使得观测变量上的复杂分布可以通过观测变量与隐变量组成的扩展空间上的联合概率分布表示，隐变量的引入使得复杂的概率分布可以由简单的分量组成。记隐变量为 Z，则观测变量与隐变量组成的扩展空间上的观测为 (X, Z)，似然函数重写为：

$$\ell(\theta) = \log p(X \mid \theta) = log \sum_{z^i} p(X, z^i \mid \theta) \tag{2}$$

即将复杂概率分布 $p(X \mid \theta)$ 写成联合概率分布 $p(X, z^i \mid \theta)$ 的和。

举个例子说明上述情况，我们从一群人中抽取了一批身高数据（X），根据该组数据估算男女身高分布（θ），利用最大似然法，只需最大化 $\log p(X \mid \theta)$ 即可，但是该分布有些复杂，很难直接计算。这时我们引入隐变量（Z），隐变量代表抽取的身高数据来自男或者女。该问题的似然函数为：

$$\begin{aligned} \ell(\theta) &= \log p(X \mid \theta) \\ &= \log \sum_{z^i} p(X, z^i \mid \theta) \\ &= \log[p(X, z^{男} \mid \theta) + p(X, z^{女} \mid \theta)] \end{aligned} \tag{3}$$

联合概率分布 $p(X, z^{男} \mid \theta)$ 和 $p(X, z^{女} \mid \theta)$ 可以认为是高斯分布，易于计算。

针对这类问题，我们的目标函数为：

$$\ell(\theta) = \log \sum_{z^i} p(X, z^i \mid \theta) \tag{4}$$

*https://github.com/leeliang/

由于 Z 是不知道，上式的联合概率分布 $p(X, z^i \mid \theta)$ 也很难计算，因此需要一些特殊的技巧。这个技巧就是找到 $p(X, z^i \mid \theta)$ 的下界函数，并最大化这个下界函数近似最大化 $\ell(\theta)$。这个技巧的数学基础就是 Jensen 不等式。

定理 1 (Jensen 不等式) 如果 $f(x)$ 是凹函数，$f[E(x)] \geq E[f(x)]$；当且仅当 $x = E(x)$ 时等式成立。

下面推导下界函数。

假设 Q 为 Z 的分布，则 $\sum\limits_{z^i} Q(z^i) = 1, Q(z^i) > 0$。(4) 式重写为：

$$\begin{aligned} \ell(\theta) &= \log \sum_{z^i} p(X, z^i \mid \theta) \\ &= \log \sum_{z^i} Q(z^i) \frac{p(X, z^i \mid \theta)}{Q(z^i)} \end{aligned} \tag{5}$$

上式中 $\sum\limits_{z^i} Q(z^i) \frac{p(X, z^i|\theta)}{Q(z^i)}$ 可以看成是 $\frac{p(X, z^i|\theta)}{Q(z^i)}$ 在分布 $Q(z^i)$ 下的期望，而对数函数又是凹函数，根据 Jensen 不等式，有：

$$\begin{aligned} \ell(\theta) &= \log \sum_{z^i} Q(z^i) \frac{p(X, z^i \mid \theta)}{Q(z^i)} \\ &\geq \sum_{z^i} Q(z^i) \log \frac{p(X, z^i \mid \theta)}{Q(z^i)} = \mathcal{L}(Q, \theta) \end{aligned} \tag{6}$$

上式的下界函数对应着两个变量 Q, θ，我们的目标函数的变量是 θ。对于特定的 θ^j，我们希望采用的下界函数在该点上能使上式成立，该条件可以帮我们确定 Q。也就是说，在 θ 空间的每一点上，都有一个下界函数，并且在该点上与目标函数的值相同。这样，我们采用一组下界函数，每个下界函数贡献一个点组成一个关于 θ 的函数，该函数等价于目标函数。但是实际处理中，我们不需要同时求出每个点对应的下界函数，我们采用的是一个迭代过程，通过下界函数不断地最大化，来使得 $\ell(\theta^j)$ 不断的提高，从而达到 $\ell(\theta)$ 的最大值，也就是 EM 算法。

如图 (1) 所示，固定 θ^{old}，根据等式成立条件计算 Q，然后固定 Q（即固定下界函数，图 (1) 中的蓝线），计算 θ^{new} 使得下界函数达到最大值，然后进入迭代直至收敛到似然函数的最大值处的 θ^*。至此，得到我们的解 $\theta^* = \arg\max\limits_{\theta} \ell(\theta)$。

刚才提到 Q 的计算原理，下面补充其具体计算过程。等式成立的条件为：

$$\frac{p(x^i, z^i \mid \theta)}{Q(z^i)} = E[\frac{p(x^i, z^i \mid \theta)}{Q(z^i)}] \tag{7}$$

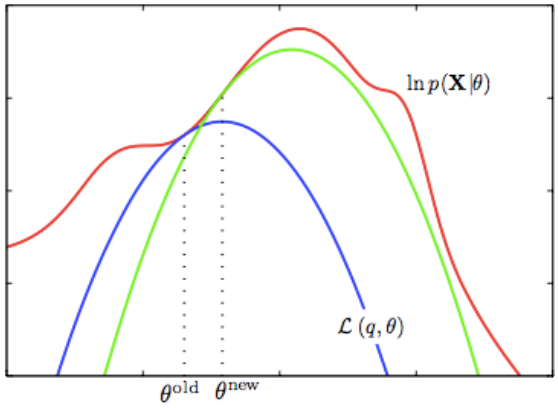

图 1: EM 算法迭代示意图

令 $\frac{p(x^i,z^i|\theta)}{Q(z^i)}$ 等于常数满足条件。则，

$$Q(z^i) = c * p(x^i, z^i \mid \theta) \tag{8}$$

c 为常数，由于 $\sum_{z^i} Q(z^i) = 1$，可得：

$$Q(z^i) = \frac{p(x^i, z^i \mid \theta)}{p(x^i \mid \theta)} \tag{9}$$

根据贝叶斯公式，可得：

$$Q(z^i) = \frac{p(x^i, z^i \mid \theta)}{p(x^i \mid \theta)} = p(z^i \mid x^i, \theta) \tag{10}$$

最后，再来看下下界函数。

$$\mathcal{L}(Q, \theta) = \sum_{z^i} Q(z^i) \log \frac{p(X, z^i \mid \theta)}{Q(z^i)} \tag{11}$$

由于 $Q(z^i)$ 在最大化 $\mathcal{L}(Q,\theta)$ 的过程中已经被固定了，最大化下界函数等价于：

$$\begin{aligned} \theta^{new} &= \arg\max_{\theta} \sum_{z^i} Q(z^i) \log \frac{p(X, z^i \mid \theta)}{Q(z^i)} \\ &= \arg\max_{\theta} \sum_{z^i} Q(z^i) \log p(X, z^i \mid \theta) \\ &= \arg\max_{\theta} \sum_{z^i} p(z^i \mid X, \theta) \log p(X, z^i \mid \theta) \\ &= \arg\max_{\theta} E_{z|X,\theta}\ \ell(\theta \mid X, Z) \end{aligned} \tag{12}$$

在不同的文献中，下界函数的写法可能有所区别，但都是等价的。还有一个问题就是迭代收敛问题，这里不记录了。

总结一下 EM 算法：

- E 步 固定 θ 推断 Q，计算似然函数 $\ell(\theta \mid X, Z)$ 关于 Z 的期望；

$$\mathcal{L}(\theta \mid \theta^{old}) = E_{z|X,\theta}\ \ell(\theta \mid X, Z)$$

- M 步 最大化下界函数求取 θ^{new};

$$\theta^{new} = \arg\max_{\theta} E_{z|X,\theta}\ \ell(\theta \mid X, Z)$$