高等数理统计 茆诗松

EM算法是一种迭代算法，最初由Dempster等提出，主要用来求后验分布的众数（即极大似然估计），

它的每一次迭代由两步组成：E步（求期望）各M步（极大化）。

一般地，以$P(\theta|Y)$表示θ的基于观测数据的后验分布密度函数，称为观测后验分布，

$p(Z|\theta,Y)$表示在给定θ和观测数据Y下潜在数据Z条件分布密度函数。

我们的目的是计算观测后验分布$p(\theta|Y)$的众数，于是，EM算法如下进行。

记$\theta^{(i)}$为第$i+1$次迭代开始时后验众数的估计值，则第i+1次迭代的两步为

E步：将$p(\theta|Y,Z)$或$\log p(\theta|Y,Z)$关于Z的条件分布求期望，从而把Z积掉，即：

$$\begin{aligned} Q(\theta|\theta^{(i)},Y) &\triangleq E_z[\log p(\theta|Y,Z)|\theta^{(i)},Y] \\ &= \int \log[p(\theta|Y,Z)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \end{aligned} \tag{1}$$

M步：将$Q(\theta|\theta^{(i)},Y)$极大化，即找一个点$\theta^{(i+1)}$，使

$$Q(\theta^{(i+1)}|\theta^{(i)},Y)=\max_{\theta} Q(\theta|\theta^{(i)},Y) \tag{2}$$

如此形成一次迭代$\theta^{(i)}\to\theta^{(i+1)}$,将上述E步和M步进行迭代直至$\|\theta^{(i+1)}-\theta^{(i)}\|$或

$\|Q(\theta^{(i+1)}|\theta^{(i)},Y)-Q(\theta^{(i)}|\theta^{(i)},Y|)\|$充分小时停止。

下面我们看一个例子。

例6.13 假设一次试验可能有四个结果，其发生的概率分别为

$\frac{1}{2}+\frac{\theta}{4},\frac{1}{4}(1-\theta),\frac{1}{4}(1-\theta),\frac{\theta}{4}$,其中$\theta\in(0,1)$

现进行了197次试验，四种结果的发生次数分别为125,18,20,34,此处观测数据为

$$Y=(y_1,y_2,y_3,y_4)=(125,18,20,34)$$

取θ的先验分布$\pi(\theta)$为平坦分布（此处即(0,1)上的均匀分布），则θ的观测后验分布为

$$\begin{aligned} p(\theta|Y) &\propto \pi(\theta)p(Y|\theta) \\ &= \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left[\frac{1}{4}(1-\theta)\right]^{y_2}\left[\frac{1}{4}(1-\theta)\right]^{y_3}\left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4} \end{aligned} \tag{3}$$

观假设第一种结果可以分解成两部分，其发生概率分别为$\frac{1}{2}$和$\frac{\theta}{4}$,

令Z和y_1-Z分别表示试验中结果落入这两部分的次数(Z是不能观测的潜在数据)，

则θ的添加后验分布为

$$\begin{aligned} p(\theta|Y,Z) &\propto \pi(\theta)p(Y,Z|\theta) \\ &= \left(\frac{1}{2}\right)^{Z}\left(\frac{\theta}{4}\right)^{y_1-Z}\left[\frac{1}{4}(1-\theta)\right]^{y_2}\left[\frac{1}{4}(1-\theta)\right]^{y_3}\left(\frac{\theta}{4}\right)^{y_4} \\ &\propto \theta^{y_1-Z+y_4}(1-\theta)^{y_2+y_3} \end{aligned} \tag{4}$$

用(3)求θ的后验众数比较麻烦，而用(4)式求后验众数则十分简单，

因此，我们可用EM算法求(3)的后验众数。

在第i次迭代中，假设有估计值$\theta^{(i)}$，则可通过E步和M步得到θ的一个新的估计。在E步中，由(1)有

$$\begin{aligned} Q(\theta|\theta^{(i)},Y) &= E^Z[(y_1-Z+y_4)\log\theta+(y_2+y_3)\log(1-\theta)|\theta^{(i)},Y] \\ &= [y_1-E^Z(Z|\theta^{(i)},Y)+y_4]\log\theta+(y_2+y_3)\log(1-\theta) \end{aligned}$$

因在$\theta^{(i)}$和Y给定下，$Z\sim b\left(y_1,\frac{2}{\theta^{(i)}+2}\right)$,故

$$E^Z(Z|\theta^{(i)},Y)=\frac{2y_1}{\theta^{(i)}+2}$$

在M步中将$Q(\theta|\theta^{(i)},Y)$对θ求导并令其为0,有

$$\begin{aligned} \theta^{(i+1)} &= \frac{y_1+y_4-E^Z(Z|\theta^{(i)},Y)}{y_1+y_2+y_3+y_4-E^Z(Z|\theta^{(i)},Y)} \\ &= \frac{\theta^{(i)}y_1+(\theta^{(i)}+2)y_4}{\theta^{(i)}y_1+(\theta^{(i)}+2)(y_2+y_3+y_4)} \\ &= \frac{159\theta^{(i)}+68}{1978^{(i)}+144} \end{aligned}$$

从$\theta^{(0)}=0.5$开始，经过四次迭代，EM算法收敛到观测后验分布的众数0.6268。

也可直接计算，若收敛，则有

$$\hat{\theta}=\frac{159\hat{\theta}+68}{197\hat{\theta}+144}$$

解得$\hat{\theta}=0.626821497$。

EM算法的最大优点是简单和稳定。

EM算法的主要目的是提供一个简单的迭代算法来计算后验众数（或MLE），人们自然会问，

如此建立的EM算法能否达到预期要求，就是说，由EM算法得到的估计序列是否收敛，

如果收敛其结果是否是$p(\theta|Y)$的最大值或局部最大值。为此，我们结出下述两个定理，

记EM算法得到的估计序列为$\theta^{(i)},i=1,2,\cdots,L(\theta|Y)=\log p(\theta|Y)$。

Theorem 1.

EM算法在每一次迭代后均提高（观测）后验密度函数值，即

$$p(\theta^{(i+1)}|Y)\geqslant p(\theta^{(i)}|Y)$$

Proof.

由全概率公式

$$p(\theta,Z|Y)=p(Z|\theta,Y)p(\theta|Y)=p(\theta|Y,Z)p(Z|Y)$$

将上式后面两项取对数有

$$\log p(\theta|Y)=\log p(\theta|Y,Z)-\log p(Z|\theta,Y)+\log p(Z|Y)$$

设现有估计$\theta^{(i)}$,将上式对Z关于$p(Z|\theta^{(i)},Y)$求期望，有

$$\begin{aligned} \log p(\theta|Y) &= \int[\log p(\theta|Y,Z)-\log p(Z|\theta,Y)+\log p(Z|Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \\ &\triangleq Q(\theta|\theta^{(i)},Y)-H(\theta|\theta^{(i)},Y)+K(\theta^{(i)},Y) \end{aligned} \tag{5}$$

其中$Q(\theta|\theta^{(i)},Y)$已在(1)中定义，

$$\begin{aligned} H(\theta|\theta^{(i)},Y) &= \int \log[p(Z|\theta,Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \\ K(\theta^{(i)},Y) &= \int \log[p(Z|Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \end{aligned}$$

分别在(5)中取θ为$\theta^{(i)}$和$\theta^{(i+1)}$并相减，有

$$\begin{aligned} &\log p(\theta^{(i+1)}|Y) - \log p(\theta^{(i)}|Y) \\ =\ &[Q(\theta^{(i+1)}|\theta^{(i)},Y) - Q(\theta^{(i)}|\theta^{(i)},Y)] - [H(\theta^{(i+1)}|\theta^{(i)},Y) - H(\theta^{(i)}|\theta^{(i)},Y)] \end{aligned}$$

由Jensen不等式

$$E^{Z|\theta^{(i)},Y}\log\left(\frac{p(Z|\theta^{(i+1)},Y)}{p(Z|\theta^{(i)},Y)}\right) \leqslant l\log\left\{E^{Z|\theta^{(i)},Y}\left(\frac{p(Z|\theta^{(i+1)},Y)}{p(Z|\theta^{(i)},Y)}\right)\right\} = 0$$

故$H(\theta^{(i+1)}|\theta^{(i)},Y) - H(\theta^{(i)}|\theta^{(i)},Y) \leqslant 0$，而$\theta^{(i+1)}$是使$Q(\theta|\theta^{(i)},Y)$达到最大的，显然

$$Q(\theta^{(i+1)}|\theta^{(i)},Y) - Q(\theta^{(i)},Y) \geqslant 0$$

□

Theorem 2.

(1)如果$p(\theta|Y)$有上界，则$L(\theta^{(i)}|Y)$收敛到某个L^*;

(2)如果$Q(\theta|\varphi)$关于θ和φ都连续，则在关于L的很一般的条件下，

由EM算法得到的估计序列θ^*是L的稳定点。

Proof.

（1）的证明由单调收敛定理立得。（2）的证明见文献[13] □

Note 3.

另一种证明思路

$$\begin{aligned} p(\theta|Y)p(Z|Y,\theta) &= p(\theta|Y,Z)p(Z|Y) \\ \ln p(\theta|Y) &= \ln[p(\theta|Y,Z)p(Z|Y)] - \ln p(Z|Y,\theta) \end{aligned}$$

$\theta = \theta^{(i)}$时，取$Z \sim p(Z|Y,\theta^{(i)})$,

$$\begin{aligned} \ln p(\theta|Y) &= E_Z\ln[p(\theta|Y,Z)p(Z|Y)] - E_Z\ln p(Z|Y,\theta) \\ &= E_Z\ln\frac{p(\theta|Y,Z)p(Z|Y)}{p(Z|Y,\theta^{(i)})} - E_Z\ln\frac{p(Z|Y,\theta)}{p(Z|Y,\theta^{(i)})} \\ &= L(\theta) + K(\theta) \\ Q(\theta) &= E_Z\ln p(\theta|Y,Z) \\ L(\theta) &= E_Z\ln\frac{p(\theta|Y,Z)p(Z|Y)}{p(Z|Y,\theta^{(i)})} \\ &= Q(\theta) + \mathrm{const} \\ K(\theta) &= -E_Z\ln\frac{p(Z|Y,\theta)}{p(Z|Y,\theta^{(i)})} \end{aligned}$$

且有

$$\begin{aligned}\ln p(\theta^{(i)}|Y) &= L(\theta^{(i)})\\ K(\theta) &\geqslant -E_Z\left(\frac{p(Z|Y,\theta)}{p(Z|Y,\theta^{(i)})}-1\right)\\ &\geqslant 0\\ \ln p(\theta|Y) &\geqslant L(\theta)\end{aligned}$$

当取

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta)$$

有

$$\ln p(\theta|Y)\geqslant L(\theta)\geqslant L(\theta^{(i)})$$

由于 $p(\theta|Y) \propto \pi(\theta)p(Y|\theta)$ ，因此对$p(\theta|Y)$的优化可以通过优化$\pi(\theta)p(Y|\theta)$进行

$$\begin{aligned}p(Y|\theta)p(Z|Y,\theta) &= p(Z,Y|\theta)\\ \pi(\theta)p(Y|\theta)p(Z|Y,\theta) &= \pi(\theta)p(Z,Y|\theta)\\ \ln[\pi(\theta)p(Y|\theta)] &= \ln[\pi(\theta)p(Z,Y|\theta)]-\ln p(Z|Y,\theta)\end{aligned}$$

E步

假设$\theta=\theta^{(i)}$，$Z\sim p(Z|Y,\theta^{(i)})$，得

$$Q(\theta) = E_Z\{\ln[\pi(\theta)p(Z,Y|\theta)]\}$$

M步

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta)$$

证明思路如下

$$\begin{aligned}\ln[\pi(\theta)p(Y|\theta)] &= E_Z\ln[\pi(\theta)p(Z,Y|\theta)]-E_Z\ln p(Z|Y,\theta)\\ &= E_Z\ln\frac{\pi(\theta)p(Z,Y|\theta)}{p(Z|Y,\theta^{(i)})}-E_Z\ln\frac{p(Z|Y,\theta)}{p(Z|Y,\theta^{(i)})}\\ &= L(\theta)-K(\theta)\\ L(\theta) &= E_Z\ln\frac{\pi(\theta)p(Z,Y|\theta)}{p(Z|Y,\theta^{(i)})}\\ &= Q(\theta)+\mathrm{const}\\ K(\theta) &= -E_Z\ln\frac{p(Z|Y,\theta)}{p(Z|Y,\theta^{(i)})}\\ \ln[\pi(\theta^{(i)})p(Y|\theta^{(i)})] &= L(\theta^{(i)})\end{aligned}$$

由于 $\ln(x)\leqslant x-1, K(\theta)\geqslant 0$，因此当$L(\theta)>L(\theta^{(i)})$时，有

$$\ln[\pi(\theta)p(Y|\theta)] \geqslant L(\theta)>L(\theta^{(i)})$$