一、为什么需要EM算法

概率模型有时既含有观测变量，又含有隐变量或者潜在变量。如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计，或者贝叶斯估计法估计模型参数。但是，当模型含有隐变量时，就不能简单地使用这些估计方法。EM算法就是含有隐变量的概率模型参数的极大似然估计法。

EM算法的每次迭代由两步组成：E步，求期望；M步，求极大。所以算法称为期望极大算法，简称EM算法。

二、EM算法的例子

《统计学习方法》例子9.1（三硬币模型）：

假设又三枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是 π，p和q。进行如下掷硬币实验：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后掷选出的硬币，掷硬币的结果，出现正面记作1，出现反面记作0；独立地重复n次实验（这里，n=10），观察结果如下：

1，1，0，1，0，0，1，0，1，1

假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三枚硬币正面出现的概率，即三枚硬币模型的参数

解：对每一次实验可以进行如下建模

前提须知公式：

$$P(A)=\sum_{B}P(A,B)$$
$$P(A,B)=P(A)\cdot P(A|B)$$

求解公式：

$$\begin{aligned}P(y|\theta)&=\sum_{z}P(y,z|\theta)=\sum_{z}P(z|\theta)P(y|z,\theta)\\&=P(z=1|\theta)P(y|z=1,\theta)+P(z=0|\theta)P(y|z=0,\theta)\\&=\begin{cases}\pi p+(1-\pi)q, & if\ \ y=1;\\ \pi(1-p)+(1-\pi)(1-q), & if\ y=0;\end{cases}\\&=\ \pi p^{y}(1-p)^{1-y}+(1-\pi)q^{y}(1-q)^{1-y}\end{aligned}$$

其中，随机变量y是观测变量，表示一次实验观测的结果是1或0；随机变量Z是隐变量，表示未观测到的掷硬币A的结果；$\theta=(\pi,p,q)$是模型参数。

将观测数据表示为$Y=(Y_1,Y_2,\dots,Y_n)^T$, 未观测数据表示为$Z=(Z_1,Z_2,\dots,Z_n)^T$则观测数据的似然函数为：

$$P(Y|\theta)=\sum_{Z}P(Z|\theta)P(Y|Z,\theta)=\prod_{j=1}^{n}P\left(y_j|\theta\right)$$

$$=\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}]$$

考虑求模型参数$\theta=(\pi,p,q)$的极大似然估计，即使用对数似然函数来进行参数估计可得：

$$\begin{aligned}\hat{\theta}&=\arg\max lnP(Y|\theta)\\&=\arg\max ln\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}]\end{aligned}$$

$$= \arg\max \sum_{j=1}^{n} ln[\pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j}]$$

三、EM 算法的导出

3.1 Jensen（琴生不等式）

若 f 是凸函数，则：

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

其中，$t \epsilon [0,1]$。同理，若 f 是凹函数，则只需将上式中的≤换成≥即可。

将上式中的 t 推广到 n 个同样成立，也即：

$$f(t_1x_1 + t_2x_2 + \cdots + t_nx_n) \leq t_1f(x_1) + t_2f(x_2) + \cdots + t_nf(x_n)$$

其中，$t_1, t_2, \ldots, t_n \epsilon [0,1]$，$t_1 + t_2 + \cdots + t_n = 1$。在概率论中常以以下形式出现：

$$\varphi(E[X]) \leq E[\varphi(X)]$$

其中 X 是随机变量，φ是凸函数，$E[X]$表示 X 的期望。

3.2 EM 算法的推导

我们面对一个含有隐变量的概率模型，目标是极大化观测数据 Y 关于参数θ的对数似然函数，即极大化：

$$L(\theta) = lnP(Y|\theta) = ln\sum_Z P(Y|Z,\theta)P(Z|\theta)$$

注意到这一极大化的主要困难是上式中有未观测数据 Z 并有包含和（Z 为离散型时）或者积分（Z 为连续型时）的对数。EM 算法采用的是通过迭代逐步近似极大化$L(\theta)$：假设在第i次迭代后θ的估计值是$\theta^{(i)}$，我们希望新的估计值θ能使$L(\theta)$增加，即$L(\theta) > L(\theta^{(i)})$并逐步达到极大值。为此，我们考虑两者的差：

$$L(\theta) - L(\theta^{(i)}) = ln\sum_Z P(Y|Z,\theta)P(Z|\theta) - lnP(Y|\theta^{(i)})$$

$$= ln\sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - lnP(Y|\theta^{(i)})$$

套用琴生不等式可得：

$$\geq \sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - lnP(Y|\theta^{(i)})$$

$$= \sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - 1 \cdot lnP(Y|\theta^{(i)})$$

$$= \sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - \sum_Z P(Z|Y,\theta^{(i)})lnP(Y|\theta^{(i)})$$

$$= \sum_Z P(Z|Y,\theta^{(i)})\left(ln\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - lnP(Y|\theta^{(i)})\right)$$

$$= \sum_Z P(Z|Y,\theta^{(i)})ln\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$$

$$L(\theta) - L(\theta^{(i)}) = \sum_Z P(Z|Y,\theta^{(i)})ln\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$$

令

$$B(\theta,\theta^{(i)}) = L(\theta^{(i)}) + \sum_Z P(Z|Y,\theta^{(i)})ln\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$$

$$L(\theta) \geq B\left(\theta, \theta^{(i)}\right)$$

即函数$B\left(\theta, \theta^{(i)}\right)$是$L(\theta)$的一个下界，此时若设$\theta^{(i+1)}$使得$B\left(\theta, \theta^{(i)}\right)$达到极大，也即

$$B\left(\theta^{(i+1)}, \theta^{(i)}\right) \geq B\left(\theta^{(i)}, \theta^{(i)}\right)$$

$$L\left(\theta^{(i+1)}\right) \geq \theta^{(i)}$$

因此，任何可以使$B\left(\theta, \theta^{(i)}\right)$增大的$\theta$也可以使$L(\theta)$增大，由于问题转化为了求解能使得$B\left(\theta, \theta^{(i)}\right)$达到极大的$\theta^{(i+1)}$，即

$$\theta^{(i+1)} = argmaxB\left(\theta, \theta^{(i)}\right)$$

$$= argmax\left(L\left(\theta^{(i)}\right) + \sum_{Z} P\left(Z \middle| Y, \theta^{(i)}\right) ln \frac{P(Y|Z,\theta)P(Z|\theta)}{P\left(Z \middle| Y, \theta^{(i)}\right) P\left(Y \middle| \theta^{(i)}\right)}\right)$$

$$= argmax\left(\sum_{Z} P\left(Z \middle| Y, \theta^{(i)}\right) ln\left(P(Y|Z,\theta)P(Z|\theta)\right)\right)$$

$$= argmax\left(\sum_{Z} P\left(Z \middle| Y, \theta^{(i)}\right) lnP(Y|Z,\theta)\right)$$

$$= argmaxQ\left(\theta, \theta^{(i)}\right)$$

四、EM 算法求解例子

使用 EM 算法求解三硬币模型