EM算法

三硬币问题：假设有A、B、C三枚硬币，抛硬币出现正面的概率分别为π, p, q。使用三枚硬币进行如下试验：首先抛掷硬币A，根据其结果来选择硬币B或者C，假设正面选B，反面选C；然后记录硬币结果，正面记为1，反面记为0。独立重复5次试验，每次试验重复抛掷B或者C 10次。问如何估计三枚硬币分别出现正面的概率。

三硬币模型可以写作：

$$P(y,\theta)=\sum_z P(y,z|\theta)=\sum_z P(z|\theta)P(y|z,\theta)$$

其中随机变量y表示观测变量，即最后观测记录的硬币结果，为1或者0；随机变量z为隐变量，表示未观测到的硬币A的抛掷结果；$\theta=(\pi,p,q)$是模型需要估计的参数。

假设观测数据记为$Y=(Y_1,Y_2,\cdots,Y_{10})^T$，未观测数据记为$z=(z_1,z_2,\cdots,z_{10})^T$，那么观测数据的似然函数为：

$$P(Y|\theta)=\sum_z P(z|\theta)P(Y|z,\theta)$$

考虑求模型$\theta=(\pi,p,q)$的极大似然估计，即求：

$$\hat{\theta}=\arg\max_\theta \log P(Y|\theta)$$

由于我们只能观察到最后的抛掷结果，至于这个结果是由硬币B抛出来的还是由硬币C抛出来的，无从知晓，所以这个过程中根据概率选择抛掷的那一枚硬币就是一个隐变量。因此我们需要用EM算法来进行求解。

E步：先初始化硬币B和C出现正面的概率为$\hat{\theta}_B^{(0)}=0.6$和$\hat{\theta}_C^{(0)}=0.5$，估计每次试验中选择B或C的概率（即硬币A是正面还是反面的概率），例如选B的概率：

$$P(z=B|y_1,\theta)=\frac{P(z=B,y_1|\theta)}{P(z=B,y_1|\theta)+P(z=C,y_1|\theta)}$$

第一次试验为5次正面5次反面

$$=\frac{(0.6)^5\times(0.4)^5}{(0.6)^5\times(0.4)^5+(0.5)^{10}}=0.45$$

相应地，选择C的概率为 1-0.45 = 0.55

计算出每次试验选择B和C的概率，然后根据试验数据进行加权求和。

M步：更新模型参数的估计值，先写出Q函数：

$$Q(\theta,\theta^{(i)}) = \sum_{j=1}^{5}\sum_{z} P(z|y_j,\theta^{(i)}) \log P(z|y_j,\theta)$$

u_j 是上面计算的隐状态

$$= \sum_{j=1}^{5} u_j \log\left(\theta_B^{y_j}(1-\theta_B)^{1-y_j}\right) + (1-u_j)\log\left(\theta_C^{y_j}(1-\theta_C)^{1-y_j}\right)$$

对上式求导并令其为零，可得第一次迭代后的参数估计结果。然后重复迭代。