EM算法

对于只有观测变量的问题，直接根据样本值通过极大似然估计的方法求解分布的参数。但是对于含有隐变量的问题，则要通过EM算法来逼近分布的参数。

1. 一个抛硬币的例子

1.1 问题描述

输入：观测变量数据Y，隐变量数据Z，联合分布$P(Y,Z|\theta)$，隐变量的条件分布$P(Z|Y,\theta)$

输出：模型参数θ

举个例子，上面的各个变量和概率，用这个抛硬币的例子来说明。

假设有3枚硬币，分别记做A, B, C。这些硬币出现正面的概率分别是π，p和q。进行以下实验：先抛硬币A，根据结果来选择硬币B或是C。假如A抛得正面，则选择硬币B，否则选择硬币C。根据选择的硬币进行抛掷，如果为正面，则记为1，反面记为0。独立重复n次实验。

假设只能看到第二阶段抛硬币（即B或C）的结果，而不能观测抛硬币的过程（选B还是C）。则在整个抛硬币过程结束后，观测结果y的概率。

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

其中$P(y|\theta)$表示给定分布的参数θ，硬币观测结果y的概率。

这个式子从统计的角度来理解，可以认为强行把隐变量z暴露出来，再对z求边缘概率，进而积分掉。右边的等式利用全概率公式写出给定z，y的条件概率$P(y|z,\theta)$

$\sum_z P(y,z|\theta)$：$P(y,z|\theta)$是y和z的联合概率。表示给定分布的参数θ，同时出现隐变量z和观测变量y的概率。（例子中表示具体选择了某一个硬币，同时利用这个硬币进行抛掷，得到观测结果y的概率）

$\sum_z P(z|\theta)$：$P(z|\theta)$是以θ为参数，z的概率分布。z的取值范围是$\{1,0\}$，分别代表正面和反面。正面$P(z=1|\theta)=\pi$，反面$P(z=0|\theta)=1-\pi$。然后根据正反选择硬币。在给定A硬币朝向的条件下，B或C硬币正面朝上的概率分别为p或q。$p=P(y=1|z=1,\theta)$，$q=P(y=1|z=0,\theta)$。则B或C硬币反面朝上的概率分别为$1-p$或$1-q$。在不知道y的结果时，把y的两种情况写在一起的话可以表示成$p^y(1-p)^{1-y}$或$q^y(1-q)^{1-y}$

因此，

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

所以$P(y|\theta)$可以表示为

$$P(y|\theta) = \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y}$$

如果独立进行n次上述抛硬币过程，用$Y=(Y^{(1)},Y^{(2)},\cdots,Y^{(n)})^T$表示观测到的数据结果，用$Y=(Z^{(1)},Z^{(2)},\cdots,Z^{(n)})^T$表示未观测数据的结果，则观测数据的似然函数可以写成

$$\begin{aligned} P(Y|\theta) &= \sum_Z P(Z|\theta)P(Y|Z,\theta) \\ &= \prod_{j=1}^{n} [\pi p^{y^{(j)}}(1-p)^{1-y^{(j)}} + (1-\pi)q^{y^{(j)}}(1-q)^{1-y^{(j)}}] \end{aligned}$$

通过极大似然估计，求解该模型的参数$\theta=(\pi,p,q)$，即

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

一般求解极大似然估计的问题，是采用对参数求偏导数，令其等于0的方式求解，而这里的问题含有隐变量Z，无法通过这种解析的方法来求解，因此采用EM算法。

1.2 用EM算法求解模型参数θ的步骤

初始化模型的参数，用$\theta_0=(\pi_0,p_0,q_0)$表示。利用EM算法对参数进行迭代更新，第i次迭代后得到的参数记做$\theta_i=(\pi_i,p_i,q_i)$。在进行第$i+1$次迭代时，过程如下。

E step:

计算在第i次模型参数$\theta_i=(\pi_i,p_i,q_i)$控制下，独立进行第j次实验观测的$y^{(j)}$是来自抛硬币B的概率是

$$\begin{aligned} \mu_{i+1} &= P(z=1|y,\theta_i) \\ &= \frac{P(z=1,y|\theta_i)}{P(y|\theta_i)} \\ &= \frac{P(z=1|\theta_i)P(y|z=1,\theta_i)}{\sum_z P(z|\theta_i)P(y|z,\theta_i)} \\ &= \frac{\pi_i(p_i)^{y^{(j)}}(1-p_i)^{1-y^{(j)}}}{\pi_i(p_i)^{y^{(j)}}(1-p_i)^{1-y^{(j)}} + (1-\pi_i)(q_i)^{y^{(j)}}(1-q_i)^{1-y^{(j)}}} \end{aligned}$$

注意分子表述的过程是先抛硬币A，有π_i的概率能够获得正面，进而选择硬币B。抛硬币B得到观测结果y。由于μ是概率，和为1,因此需要分母作归一化。

M step 更新模型参数

$$\pi_{i+1} = \frac{1}{n}\sum_{j=1}^{n}\mu_{i+1}^{(j)}$$

$$p_{i+1} = \frac{\sum_{j=1}^{n} u_{i+1}^{(j)} y^{(j)}}{\sum_{j=1}^{n} u_{i+1}^{(j)}}$$

$$q_{i+1} = \frac{\sum_{j=1}^{n} (1-u_{i+1}^{(j)}) y^{(j)}}{\sum_{j=1}^{n} (1-u_{i+1}^{(j)})}$$

从感性的认识上来看，第一个式子中，把独立的n次试验，得到的实验观测的$y^{(j)}$是来自抛硬币B的概率的平均值，则为抛硬币A后会选择硬币B的概率π。

第二个式子，可以理解为在观测到了最终的结果$y^{(j)}$中，有μ_{i+1}的部分来自于硬币B贡献的，因此分子要用μ_{i+1}去对$y^{(j)}$进行加权。由于不是百分百由B贡献，因此分母进行归一化时，除的是μ_{i+1}的和。

对于第三个式子的理解和第二个式子类似。

重复以上的E step和M step直到收敛，即为EM算法的过程。

该算法如果系统表述，则为：

输入：观测变量Y，隐变量Z，联合分布$P(Y,Z|\theta)$，隐变量的条件概率$P(Z|Y,\theta)$

输出：模型参数θ

1. 选择参数的初值θ_0（EM算法对初始化的参数敏感）
2. E步：记θ_i为第i次迭代参数θ的估计值，在第$i+1$次迭代的E步，计算Q函数
$$\begin{aligned} Q(\theta,\theta_i) &= E_Z[\log P(Y,Z|\theta)|Y,\theta_i] \\ &= \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta_i) \end{aligned}$$
这里的$P(Z|Y,\theta_i)$是在给定观测数据Y和当前的参数估计θ_i下，隐变量Z的条件概率。Q函数是关于θ的函数。
3. M步：求使得$Q(\theta,\theta_i)$极大化的θ，确定第i+1次迭代的参数估计值θ_{i+1}
4. 重复2，3步，直到收敛

Q函数的意义：联合概率的对数似然函数$\log P(Y,Z|\theta)$对隐变量条件概率$P(Z|Y,\theta_i)$的期望。因此是以$P(Z|Y,\theta_i)$为权重，对$\log P(Y,Z|\theta)$进行加权求和的过程。

1.3 EM算法的说明

1. 参数的初始值可以任意选择，但是EM算法对初始值的选择是敏感的。
2. 在E步求Q函数，Q函数中的Z是未观测数据（隐变量），Y是观测数据，在$Q(\theta,\theta_i)$中，第一个参数θ是可以调的变量，在M步中通过调节这个参数使得Q函数最大化。第二个参数θ_i是参数当前的估计，在M步中固定。
3. 在M步中求使得Q函数最大化的θ，即为θ_{i+1}，这样完成一次迭代$\theta_i \rightarrow \theta_{i+1}$

4. 迭代停止的条件，若满足

$$\|\theta_{i+1}-\theta_i\| \le \epsilon_1 \text{或} \|Q(\theta_{i+1},\theta_i)-Q(\theta_i,\theta_i)\| \le \epsilon_2$$

则停止迭代，ϵ_1和ϵ_2是超参数。

1.4 为什么EM算法长这样

这一节主要回答Q函数是怎么得到的。

根据一般的极大似然估计法的步骤，首先我们要先写出观测数据的似然函数，然后取对数，最后求极大值得到模型的参数。按照这个步骤，首先，似然函数

$$\begin{aligned} L(\theta)=P(Y|\theta)&=\sum_Z P(Y,Z|\theta) \\ &=\sum_Z P(Y|Z,\theta)P(Z|\theta) \end{aligned}$$

因为含有隐变量Z，所以把似然函数强行写成含有隐变量Z的形式。

其次，取对数

$$l(\theta)=\log L(\theta)=\log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right)$$

最后，取对数似然的极大。上文讲到了这个对数似然函数含有隐变量Z，因此无法直接求偏导取极大，而要用迭代的方法。因此要随机猜一个初始值θ_0。

不失一般性，假设迭代进行完第i轮，得到了参数θ_i时，我们希望在第$i+1$轮时得到的新估计值能够继续使对数似然函数增大，即$l(\theta)>l(\theta_i)$。

我们无法直接使得$l(\theta)$取极大，退而求其次，我们找到$l(\theta)$的下界，对下界求极大。利用Jensen不等式，

$$\begin{aligned} l(\theta)&=\log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right) \\ &=\log\left(\sum_Z P(Z|Y,\theta_i)\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta_i)}\right) \\ &\ge \sum_Z P(Z|Y,\theta_i)\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta_i)} \end{aligned}$$

（这里把$\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta_i)}$看成整体）当$\theta=\theta_i$时，不等号可以取到等号。

定义下界为关于θ的函数

$$B(\theta_i,\theta)=\sum_Z P(Z|Y,\theta_i)\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta_i)}$$

取极大值，并将常数项去掉

$$
\begin{aligned}
\arg\max_{\theta} B(\theta_i,\theta) &= \arg\max_{\theta}\sum_{Z} P(Z|Y,\theta_i)\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta_i)} \\
&= \arg\max_{\theta}\sum_{Z} P(Z|Y,\theta_i)\log P(Y|Z,\theta)P(Z|\theta) \\
&= \arg\max_{\theta}\sum_{Z} P(Z|Y,\theta_i)\log P(Y,Z|\theta) \\
&= \arg\max_{\theta} Q(\theta,\theta_i)
\end{aligned}
$$

可以看到我们对下界取极大值，就是我们在上文中提到的对Q函数取极大值的过程。记住，**Q函数是EM算法的核心。**

1.5 EM算法收敛的必然性

EM算法有以下性质：每次迭代得到的$P(Y|\theta_i)$要不小于上一轮的结果，即

$$
P(Y|\theta_{i+1}) \geq P(Y|\theta_i)
$$

证明如下

$$
P(Y|\theta) = P(Y|Z,\theta) = \frac{P(Y,Z|\theta)}{P(Z|Y,\theta)}
$$

对其取对数得到

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

已知Q函数$Q(\theta,\theta_i) = \sum_Z P(Z|Y,\theta_i)\log P(Y,Z|\theta)$。定义H函数

$$
H(\theta,\theta_i) = \sum_{Z} P(Z|Y,\theta_i)\log P(Z|Y,\theta)
$$

则

$$
\begin{aligned}
Q(\theta,\theta_i) - H(\theta,\theta_i) &= \sum_{Z} P(Z|Y,\theta_i)\log\frac{P(Y,Z|\theta)}{P(Z|Y,\theta)} \\
&= \sum_{Z} P(Z|Y,\theta_i)\log P(Y|\theta) \\
&= \log P(Y|\theta)
\end{aligned}
$$

发现对数似然函数可以由Q函数和H函数的差来表达。回到命题要求证明每次迭代得到的似然函数都会增大，利用log函数的单调性，可以将命题转化为证明对数似然函数每次迭代都会增大。因此计算

$$\log P(Y|\theta_{i+1}) - \log P(Y|\theta_i) = [Q(\theta_{i+1}, \theta_i) - Q(\theta_i, \theta_i)] - [H(\theta_{i+1}, \theta_i) - H(\theta_i, \theta_i)]$$

其中，$[Q(\theta_{i+1}, \theta_i) - Q(\theta_i, \theta_i)]$项可以利用EM算法每次迭代是对Q函数求极大值，因此 $Q(\theta_{i+1}, \theta_i) \geq Q(\theta_i, \theta_i)$。接下来证明$[H(\theta_{i+1}, \theta_i) - H(\theta_i, \theta_i)] \leq 0$。

根据Jensen不等式，有

$$\begin{aligned} H(\theta_{i+1}, \theta_i) - H(\theta_i, \theta_i) &= \sum_Z P(Z|Y, \theta_i) \log \frac{P(Z|Y, \theta_{i+1})}{P(Z|Y, \theta_i)} \\ &\leq \log \left(\sum_Z \frac{P(Z|Y, \theta_{i+1})}{P(Z|Y, \theta_i)} P(Z|Y, \theta_i) \right) \\ &= \log \left(\sum_Z P(Z|Y, \theta_{i+1}) \right) = \log 1 = 0 \end{aligned}$$

由此可得

$$\log P(Y|\theta_{i+1}) - \log P(Y|\theta_i) \geq 0$$

原命题得证。