人工智能的数学基础

华东师范大学 数学科学学院 黎芳(教授) 2019年11月11日

Chapter 9 EM算法及其推广（Expectation Maximization）

Table of Contents

9.1 EM算法的引入

EM--1977, Dempster等人提出，用于含隐变量的概率模型参数估计。

9.1.1 EM算法

例 **9.1** （三硬币模型） 假设有3 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. 进行如下掷硬币试验， 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 次试验（这里，n=10）, 观测结果如下:

$$1, 1, 0, 1, 0, 0, 1, 0, 1, 1$$

假设只能观测到掷硬币的结果，不能观测掷硬币的过程， 问如何估计三硬币正面出现的概率， 即三硬币模型的参数.

解：三硬币模型可以写作

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

这里， 随机变量y是观测变量，表示一次试验观测的结果是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．这一模型是以上数据的生成模型．注意，随机变量y的数据可以观测，随机变量z的数据不可观测。

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

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

$$P(Y|\theta)=\prod_{j=1}^{n}\left[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}\right]$$

考虑求模型参数$\theta=(\pi,p,q)$的极大似然估计

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

EM 算法

首先选取参数的初值， 记作$\theta^{(0)}=(\pi^{(0)},p^{(0)},q^{(0)})$， 第$i$次迭代参数的估计值为$\theta^{(i)}=(\pi^{(i)},p^{(i)},q^{(i)})$， EM算法的第$i+1$次迭代如下：

E步：计算在模型参数$\pi^{(i)},p^{(i)},q^{(i)}$下观测数据$y_j$来自掷硬币B的概率

$$\mu_j^{(i+1)}=\frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}}$$

M步：计算模型参数的新估计值

$$\pi^{(i+1)}=\frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)},\ p^{(i+1)}=\frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}},\ q^{[i+1]}=\frac{\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)y_j}{\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)}$$

$\pi^{(0)}=0.5,\quad p^{(0)}=0.5,\quad q^{(0)}=0.5$

$\mu_j^{(1)}=0.5,\quad j=1,2,\cdots,10$

$\pi^{(1)}=0.5,\quad p^{(1)}=0.6,\quad q^{(1)}=0.6$

$\mu_j^{(2)}=0.5,\quad j=1,2,\cdots,10$

$\pi^{(2)}=0.5,\quad p^{(2)}=0.6,\quad q^{(2)}=0.6$

$\hat{\pi}=0.5,\quad \hat{p}=0.6,\quad \hat{q}=0.6$

```
y = [1 1 0 1 0 0 1 0 1 1];
% pi = 0.5; p = 0.5; q = 0.5;
pi = 0.4; p = 0.6; q = 0.7;
for i = 1:5
    mu = pi*p.^y.*(1-p).^(1-y)./(pi*p.^y.*(1-p).^(1-y)+(1-pi)*q.^y.*(1-q).^(1-y));
    pi = mean(mu);
    p = sum(mu.*y)./sum(mu);
    q = sum((1-mu).*y)./sum(1-mu);
    [pi p q]
end
```

```
ans = 1×3
    0.4064    0.5368    0.6432
ans = 1×3
    0.4064    0.5368    0.6432
ans = 1×3
    0.4064    0.5368    0.6432
ans = 1×3
    0.4064    0.5368    0.6432
ans = 1×3
    0.4064    0.5368    0.6432
```

用Y表示观测随机变量的数据，Z表示隐随机变量的数据。Y和Z连在一起称为完全数据（complete-data），观测数据Y称为不完全数据（incomplete-data）. 假设θ是要估计的参数，那么不完全数据的似然函数为$P(Y|\theta)$，完全数据的似然函数为$P(Y,Z|\theta)$.

算法**9.1** （EM算法）

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

输出：模型参数θ

(1) 选择参数的初值$\theta^{(0)}$，开始迭代；

(2) E步：记$\theta^{(i)}$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

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}](\text{条件期望}) \\ &= \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)}) \end{aligned}$$

(3) M步：
$$\theta^{(i+1)} = \arg\max_\theta Q(\theta,\theta^{(i)})$$

(4) 重复第(2)步和第(3)步，直到收敛.

定义**9.1** （Q函数）完全数据的对数似然函数$\log P(Y,Z|\theta)$关于在给定观测数据Y和当前参数$\theta^{(i)}$下对未观测数据Z的条件概率分布$P(Z|Y,\theta^{(i)})$的期望称为Q函数，即

$$Q(\theta, \theta^{(i)}) = E_Z\{\log P(Y, Z|\theta)|Y, \theta^{(i)}]$$

下面关于EM算法作几点说明：

- 步骤(1) 参数的初值可以任意选择，但需注意EM算法对初值是敏感的.
- 步骤(2) E步求$Q(\theta, \theta^{(i)})$. Q函数式中Z是未观测数据，Y是观测数据. 注意，$Q(\theta, \theta^{(i)})$的第1个变元表示要极大化的参数，第2个变元表示参数的当前估计值。每次迭代实际在求Q函数及其极大.
- 步骤(3) M步求$Q(\theta, \theta^{(i)})$的极大化，得到$\theta^{(i+1)}$，完成一次迭代$\theta^{(i)} \to \theta^{(i+1)}$. 后面将证明每次迭代使似然函数增大或达到局部极值.
- 步骤(4) 停止条件，一般是对较小的正数$\varepsilon_1, \varepsilon_2$, 满足

$$\|\theta^{(i+1)} - \theta^{(i)}\| < \varepsilon_1 \text{ 或 } \|Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})| < \varepsilon_2.$$

9.1.2 EM算法的导出

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

事实上，EM算法是通过迭代逐步近似极大化$L(\theta)$的。 假设在第i次迭代后θ的估计值是$\theta^{(i)}$。我们希望新估计值θ能使$L(\theta)$增加，即$L(\theta) > L(\theta^{(i)})$，并逐步达到极大值。为此，考虑两者的差：

$$L(\theta) - L(\theta^{(i)}) = \log \left(\sum_Z P(Y|Z, \theta) P(Z|\theta) \right) - \log P(Y|\theta^{(i)})$$

利用Jensen不等式

$$\begin{aligned} L(\theta) - L(\theta^{(i)}) &= \log \left(\sum_Z P(Y|Z, \theta^{(i)}) \frac{P(Y|Z, \theta) P(Z|\theta)}{P(Y|Z, \theta^{(i)})} \right) - \log P(Y|\theta^{(i)}) \\ &\geq \sum_Z P(Z|Y, \theta^{(i)}) \log \frac{P(Y|Z, \theta) P(Z|\theta)}{P(Z|Y, \theta^{(i)})} - \log P(Y|\theta^{(i)}) \\ &= \sum_Z P(Z|Y, \theta^{(i)}) \log \frac{P(Y|Z, \theta) P(Z|\theta)}{P(Z|Y, \theta^{(i)}) P(Y|\theta^{(i)})} \end{aligned}$$

$$B(\theta, \theta^{(i)}) \triangleq \mathscr{L}(\theta^{(i)}) + \sum_Z P(Z|Y, \theta^{(i)}) \log \frac{P(Y|Z, \theta) P(Z|\theta)}{P(Z|Y, \theta^{(i)}) P(Y|\theta^{(i)})}$$

$$L(\theta) \geq B(\theta, \theta^{(i)}) \text{(下界)}$$

$$L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)})$$

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

$$
\begin{aligned}
\theta^{(i+1)} &= \arg\max_{\theta}\left(L(\theta^{(i)}) + \sum_{Z} P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}\right) \\
&= \arg\max_{\theta}\left(\sum_{Z} P(Z|Y,\theta^{(i)}) \log\left(P(Y|Z,\theta)P(Z|\theta)\right)\right] \\
&= \arg\max_{\theta}\left(\sum_{Z} P(Z|Y,\theta^{(i)}) \log P(Y,Z|\theta)\right) \\
&= \arg\max_{\theta} Q(\theta,\theta^{(i)})
\end{aligned}
$$

EM算法是通过不断求解下界的极大化逼近求解对数似然函数极大化的算法。

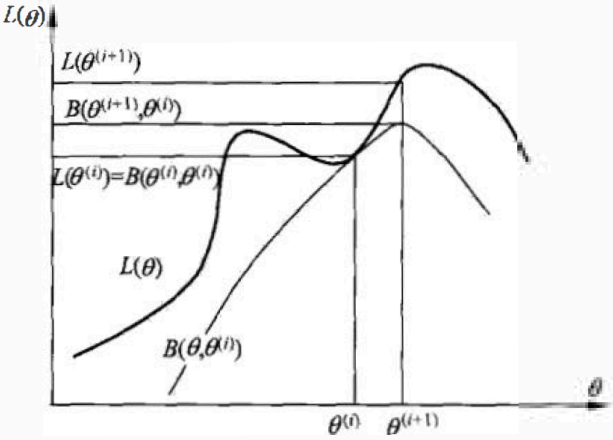


图 9.1 EM 算法的解释

9.1.3 EM算法在无监督学习中的应用

9.2 EM算法的收敛性

定理**9.1**设$P(Y|\theta)$为观测数据的似然函数，$\theta^{(i)}(i=1,2,\cdots)$为EM算法得到的参数估计序列，$P(Y|\theta^{(i)})(i=1,2,\cdots)$为对应的似然函数序列，则$P(Y|\theta^{(i)})$是单调递增的，即

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$$

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

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

$$Q(\theta,\theta^{(i)}) = \sum_{Z} \log P(Y,Z|\theta) P(Z|Y,\theta^{(i)})$$

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

于是对数似然函数可以写成

$$\log P(Y|\theta) = Q(\theta,\theta^{(i)}) - H(\theta,\theta^{(i)})$$

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

下证上式右端非负

$$Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) \geq 0$$

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

这里的不等号是由Jensen不等式得到的.

定理**9.2** 设$L(\theta) = \log P(Y|\theta)$为观测数据的对数似然函数，$\theta^{(i)}(i = 1, 2, \cdots)$为EM算法得到的参数估计序列，$L(\theta^{(i)})(i = 1, 2, \cdots)$为对应的对数似然函数序列

(1) 如果$P(Y|\theta)$有上界，则$L(\theta^{(i)}) = \log P(Y|\theta^{(i)})$收敛到某一值$L^*$.

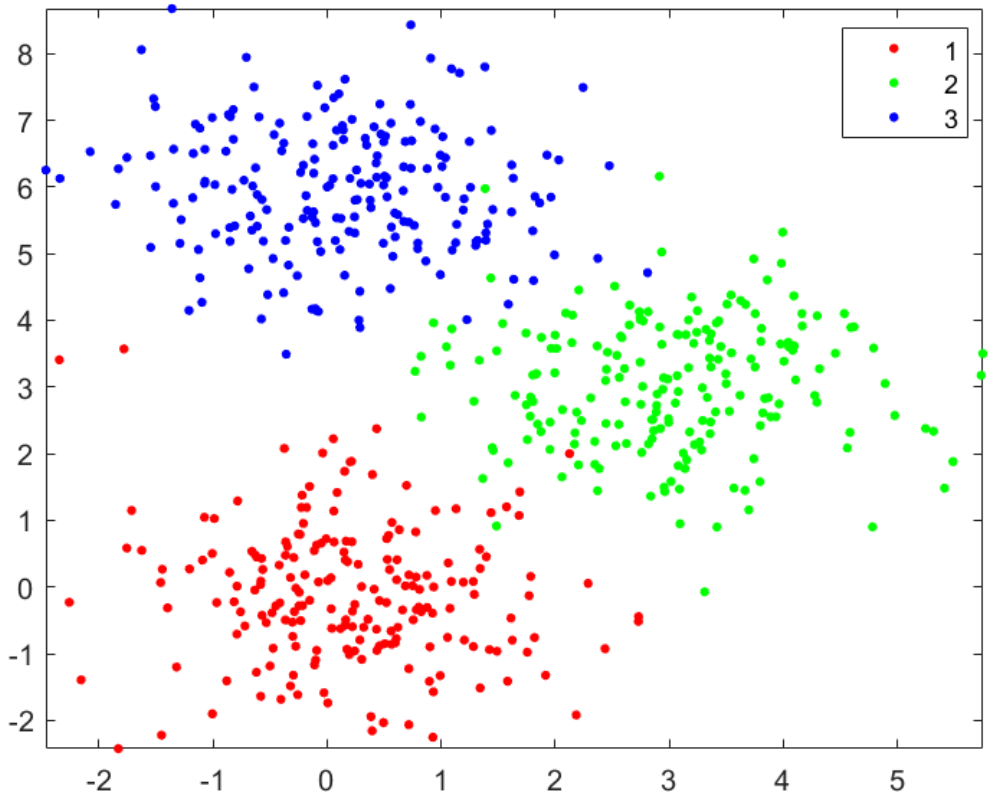
(2) 在函数$Q(\theta, \theta')$与$L(\theta)$满足一定条件下，由EM算法得到的参数估计序列$\theta^{(i)}$的收敛值θ^*是$L(\theta)$的稳定点.

定理只能保证参数估计序列收敛到对数似然函数序列的稳定点，不能保证收敛到极大值点．所以在应用中，初值的选择变得非常重要，常用的办法是选取几个不同的初值进行迭代，然后对得到的各个估计值加以比较，从中选择最好的。

9.3 EM算法在高斯混合模型学习中的应用

9.3.1 高斯混合模型

```
close all
x1 = randn(200,2)+0;
x2 = randn(200,2)+3;
x3 = randn(200,2)+[0,6];
x = [x1;x2;x3];
y = [ones(200,1); 2*ones(200,1); 3*ones(200,1)];
figure,gscatter(x(:,1),x(:,2),y)
axis([min(x(:,1)) max(x(:,1)) min(x(:,2)) max(x(:,2))])
```



定义**9.2** （高斯混合模型） 高斯混合模型是指具有如下形式的概率分布模型：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$$

其中，α_k 是系数，$\alpha_k\geq 0,\quad \sum_{i=1}^{K}\alpha_k=1;\quad \phi(y|\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\,\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

称为第k个分模型。

9.3.2 高斯混合模型参数估计的EM算法

假设观测数据$y_1,y_2,\cdots,y_N$由高斯混合模型生成，

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$$

其中$\theta=(\alpha_1,\alpha_2,\cdots,\alpha_\kappa;\theta_1,\theta_2,\cdots,\theta_\kappa)$.

1. 明确隐变量，写出完全数据的对数似然函数

可以假设观测数据y_j，$j=1,\cdots N$是这样产生的：首先依概率α_k选择第k个高斯分布$\phi(y|\theta_k)$，然后依第k个高斯分布生成观测数据，这时观测数据y_j，$j=1,\cdots N$是已知的，反映观测数据y_j来自第k个高斯分布的参数是未知的，用隐变量γ_{jk}表示，定义如下：

$$\gamma_{jk}=\begin{cases}1, \text{第}j\text{个观测数据来自第k个分模型}\\ 0, \text{否则}\end{cases}$$

$j=1,2,\cdots,N; k=1,2,\cdots,K$.

γ_{jk}是0-1随机变量.

完全数据为：$\left(y_j,\gamma_{j_1},\gamma_{j_2},\cdots,\gamma_{jK}\right)$，$j=1,2,\cdots,N$

完全数据的似然函数为：

$$\begin{aligned}P(y,\gamma|\theta) &= \prod_{j=1}^{N}P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}|\theta)\\ &=\prod_{k=1}^{K}\prod_{j=1}^{N}\left[\alpha_k\phi(y_j|\theta_k)\right]^{\gamma_{jk}}\\ &=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left(\phi(y_j|\theta_k)\right]^{\gamma_{jk}}\\ &=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\,\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}\end{aligned}$$

其中$n_k=\sum_{j=1}^{N}\gamma_{jk}$，$\sum_{k=1}^{K}n_k=N$.

完全数据的对数似然函数为

$$\log P(y,\gamma|\theta)=\sum_{k=1}^{K}\left\{n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_k\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}$$

2. EM算法的**E**步：确定**Q**函数

$$\begin{aligned}Q(\theta,\theta^{(i)}) &= E[\log P(y,\gamma|\theta)|y,\theta^{(i)}]\\ &=E\left\{\sum_{k=1}^{K}n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}\\ &=\sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E\gamma_{jk})\log\alpha_k+\sum_{j=1}^{N}(E\gamma_{jk})\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}\end{aligned}$$

记$\hat{\gamma}_{jk}=E(\gamma_{jk}|y,\theta)$

$$
\begin{aligned}
\hat{\gamma}_{jk} &= E(\gamma_{jk}|y,\theta) = P(\gamma_{jk}=1|y,\theta) \\
&= \frac{P(\gamma_{jk}=1, y_j|\theta)}{\sum_{k=1}^{K} P(\gamma_{jk}=1, y_j|\theta)} \\
&= \frac{P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)}{\sum_{k=1}^{K} P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)} \\
&= \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j|\theta_k)}, \quad j=1,2,\cdots,N; k=1,2,\cdots,K
\end{aligned}
$$

$\hat{\gamma}_{jk}$是在当前模型参数下第j个观测数据来自第k个分模型的概率， 称为分模型k对y_j的响应度。

将$\hat{\gamma}_{jk} = E\gamma_{jk}$及$n_k = \sum_{j=1}^{N} E\gamma_{jk}$代入Q的表达式，即得

$$
Q(\theta, \theta^{(i)}) = \sum_{k=1}^{K} \left\{ n_k \log \alpha_k + \sum_{j=1}^{N} \hat{\gamma}_{jk} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2}(y_j - \mu_k)^2 \right] \right\} \qquad (9.29)
$$

3.EM算法的**M**步：

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

(9.29)对各参数求导并令为0，

$$
\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \quad \hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}(y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \quad \hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, k = 1,2,\cdots,K
$$

算法9.2（高斯混合模型参数估计的**EM**算法）

输入：观测数据$y_1, y_2, \cdots, y_N$，高斯混合模型

输出：高斯混合模型参数

- (1) 取参数的初始值开始迭代
- (2) E步：依据当前模型参数，计算分模型k对观测数据y_j的响应度

$$
\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j|\theta_k)}, \quad j=1,2,\cdots,N; \quad k=1,2,\cdots,K
$$

- (3) M步：计算新一轮的迭代参数

$$\hat{\mu}_k=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}},\quad \hat{\sigma}_k^2=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}},\quad \hat{\alpha}_k=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}}{N},k=1,2,\cdots,K$$

- (4) 重复(2)(3)直至收敛.

9.4 EM算法的推广

9.4.1 F函数的极大--极大算法

9.3 定义（F函数）假设隐变量数据Z的概率分布为$\tilde{P}(Z)$，定义分布$\tilde{P}$与参数θ的函数$F(\tilde{P},\theta)$如下：

$$F(\tilde{P},\theta)=E_{\tilde{P}}[\log P(Y,Z|\theta)]+H(\tilde{P})\quad(9.33)$$

称为F函数，其中$H(\tilde{P})=-E_{\tilde{P}}\log\tilde{P}(Z)$是分布$\tilde{P}(Z)$的熵。

9.1 引理 对于固定的θ，存在唯一的分布$\tilde{P}_\theta$极大化$F(\tilde{P},\theta)$，这时$\tilde{P}_\theta$由下式给出：

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

并且$\tilde{P}_\theta$随θ连续变化。

证明：对于固定的θ,可以求得使$F(\tilde{P},\theta)$达到极大的分布$\tilde{P}_\theta(Z)$．为此，引进拉格朗日乘子λ，拉格朗日函数为

$$L=E_{\tilde{P}}\log P(Y,Z|\theta)-E_{\tilde{P}}\log\tilde{P}(Z)+\lambda\left(1-\sum_Z\tilde{P}(Z)\right)$$

$$\frac{\partial L}{\partial\tilde{P}(Z)}=\log P(Y,Z|\theta)-\log\tilde{P}(Z)-1-\lambda=0$$

$$\lambda=\log P(Y,Z|\theta)-\log\tilde{P}_\theta(Z)-1\Big]$$

$$\frac{P(Y,Z|\theta)}{\tilde{P}_\theta(Z)}=\mathrm{e}^{1+\lambda}$$ 即$P(Y,Z|\theta)$与$\tilde{P}_\theta(Z)$成比例.

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

利用约束条件 $\sum_Z \widetilde{P}_\theta(Z)=1,\ P(Y|\theta)=C \ \Rightarrow \widetilde{P}_\theta(Z)=P(Z|Y,\theta)$

引理**9.2** 若 $\widetilde{P}_\theta(Z)=P(Z|Y,\theta)$，则 $F(\widetilde{P},\theta)=\log P(Y|\theta)$.

定理**9.3** 设 $L(\theta)=\log P(Y|\theta)$ 为观测数据的对数似然函数，$\theta^{(i)}, i=1,2,\cdots,$ 为EM算法得到的参数估计序列，函数 $F(\widetilde{P},\theta)$ (9.33) $F(\widetilde{P},\theta)$ $\widetilde{P}^*$ θ^* $L(\theta)$ θ^* 由式* 定义。如果 在 和 有局部极大值，那么 也在 有局部极大值．类似地，如果 $F(\widetilde{P},\theta)$ $\widetilde{P}$ θ^* $L(\theta)$ θ^* 在 和 达到全局最大值，那么 也在 达到全局最大值。

定理**9.4** EM算法的一次迭代可由F函数的极大-极大算法实现.

设 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计，$\widetilde{P}^{(i)}$ 为第 i 次迭代 $\widetilde{P}$ 的估计。在第 i+1 次迭代的两步为：

(1) 对固定的 $\theta^{(i)}$，求 $\widetilde{P}^{(i+1)}$ 使 $F\left(\widetilde{P},\theta^{(i)}\right)$ 极大化；

(2) 对固定的 $\widetilde{P}^{(i+1)}$，求 $\theta^{(i+1)}$ 使 $F\left(\widetilde{P}^{(i+1)},\theta\right)$ 极大化。

证明：

(1) $\widetilde{P}^{(i+1)}(Z)=\widetilde{P}_{\theta^{(i)}}(Z)=P(Z|Y,\theta^{(i)})$

$$F\left(\widetilde{P}^{(i+1)},\theta\right)=E_{\widetilde{P}^{(i+1)}}[\log P(Y,Z|\theta)]+H\left(\widetilde{P}^{(i+1)}\right)$$

$$=\sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)})+H\left(\widetilde{P}^{(i+1)}\right)$$

$$F\left(\widetilde{P}^{(i+1)},\theta\right)=Q(\theta,\theta^{(i)})+H\left(\widetilde{P}^{(i+1)}\right)$$

(2) $\theta^{(i+1)}=\arg\max_\theta F\left(\widetilde{P}^{(i+1)},\theta\right)=\arg\max_\theta Q(\theta,\theta^{(i)})$.

9.4.2 GEM算法

算法**9.3**（GEM算法I）

输入：观测数据，F函数；

输出：模型参数

- (1) 初始化参数 $\theta^{(0)}$，开始迭代

- (2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，$\tilde{P}^{(i)}$ 为第 i 次迭代函数 $\tilde{P}$ 的估计值，求 $\tilde{P}^{(i+1)}$ 使 $F\left(\tilde{P}, \theta^{(i)}\right)$ 极大化；
- (3) 第 2 步：求 $\theta^{(i+1)}$ 使 $F\left(\tilde{P}^{(i+1)}, \theta\right)$ 极大化；
- (4) 重复(2)-(3)，直到收敛.

算法**9.4**（GEM算法2）

输入：观测数据，Q函数；

输出：模型参数

- (1) 初始化参数 $\theta^{(0)}$，开始迭代
- (2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \\ &= \sum_Z P(Z|Y, \theta^{(i)}) \log P(Y, Z|\theta) \end{aligned}$$

- (3) 第 2 步：求 $\theta^{(i+1)}$ 使

$$Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)})$$

- (4) 重复(2)-(3)，直到收敛.

算法**9.5**（GEM算法3）

输入：观测数据，Q函数；

输出：模型参数

- (1) 初始化参数 $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_d^{(0)}\right)$，开始迭代
- (2) 第 i+1 次迭代，第 1 步：记 $\theta^{(i)} = \left(\theta_1^{(i)}, \theta_2^{(i)}, \cdots, \theta_d^{(i)}\right)$ 为参数 $\theta = (\theta_1, \theta_2, \cdots, \theta_d)$ 的估计值，计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \\ &= \sum_Z P(Z|Y, \theta^{(i)}) \log P(Y, Z|\theta) \end{aligned}$$

- (3) 第 2 步：进行 d 次条件极大化：

首先， 在$\theta_2^{(i)}, \cdots, \theta_d^{(i)}$保持不变的条件下求使$Q(\theta, \theta^{(i)})$达到极大的$\theta_1^{(i+1)}$; 然后，
在$\theta_1 = \theta_1^{(i+1)}, \theta_j = \theta_j^{(i)}, j = 3, 4, \cdots, k$的条件下求使$Q(\theta, \theta^{(i)})$达到极大值的$\theta_2^{(i+1)}$. 如此继续，经过d次条件极大化，得到$\theta^{(i+1)} = \left(\theta_1^{(i+1)}, \theta_2^{(i+1)}, \cdots, \theta_d^{(i+1)}\right)$使得

$$Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)})$$

- (4) 重复(2)-(3)，直到收敛.

作业

习题9.3 已知观测数据-67,-48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75. 试估计两个分量的高斯混合模型的参数.