# 第9章 EM算法极其推广

# 1 EM算法的引入

# EM方法

- 记号

  观察数据(observerd data, incomplete data) $Y$

  隐藏数据(unobserved data, latent variable) $Z$

  完全数据(complete-data) $(Y, Z)$

  模型参数$\theta$【模型为某个参数$\theta$未知的概率分布】

- 给定观测数据$Y$，其概率分布是$P(Y \mid \theta)$，$Y$和$Z$的联合概率分布$P(Y, Z \mid \theta)$
  - 不完全数据$Y$的似然函数$P(Y \mid \theta)$，对数似然函数$L(\theta) = \log P(Y \mid \theta)$
  - 完全数据的对数似然函数是$\log P(Y, Z \mid \theta)$
  - 【注】似然函数 $P(Y \mid \theta)$描述参数$\theta$的条件下，出现观测值$Y$的合理性（可能性）
- 极大化$\log P(Y, Z \mid \theta)$
  - 挑战：包含了有未观测数据并有包含和(或积分)的对数

# EM算法

EM算法通过迭代方式，对期望$L(\theta)=\log P(Y,Z\mid\theta)$极大似然估计。

【注意】$P(Y,Z\mid\theta)$可用于描述出现观测值的可能性，极大似然寻到使得出现观测值可能性最大的参数

每次迭代：E步求期望；M步求极大化

【核心：迭代方式，过程单调递增，且有上界=>收敛】

【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$P(Y,Z\mid\theta)$，条件分布$P(Z\mid Y,\theta)$

输出：模型参数$\theta$

# EM算法

1)选择参数的初值$\theta^{(0)}$

2)E步

记$\theta^{(i)}$为第$i$次迭代参数$\theta$的估计值，在第$i+1$次迭代的E步，计算Q函数

$$Q\left(\theta,\theta^{(i)}\right)=E_Z\left[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}\right]=\sum\nolimits_Z \log P(Y,Z\mid\theta)P\left(Z\mid Y,\theta^{(i)}\right)$$

【注】基于隐含变量$Z$ 分布（$P\left(Z\mid Y,\theta^{(i)}\right)$，在给定观测数据$Y$和当前的参数估计$\theta^{(i)}$下，估算出来的隐变量$Z$的条件概率分布）的完全数据的对数似然函数（$\log P(Y,Z\mid\theta)$）的期望。

3)M步

求使$Q\left(\theta,\theta^{(i)}\right)$极大化的$\theta$。确定第$i+1$次迭代的参数的估计值$\theta^{(i+1)}$

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

4)重复第2)步和第3)步，直到收敛

# 三硬币模型

**例 9.1 (三硬币模型)**

假设有3枚硬币$\mathrm{A}, \mathrm{B}, \mathrm{C}$，它们正面朝上的概率分别是$\pi, p$和$q$。进行如下掷硬币试验：

先掷硬币 $\mathrm{A}$, 根据其结果选出硬币B或硬币C。$\mathrm{A}$正面选硬币B, $\mathrm{A}$ 反面选硬币C;

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 $(n = 5$ ), 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。

如何估计三硬币正面出现的概率？即三硬币模型的参数$\pi, p$和$q$。

# 三硬币模型

A, B, C 正面出现的概率分别是$\pi, p$和$q$； A正面选硬币B, A 反面选硬币C

【解】$y$是观测变量，观测值为1或0；$z$表示未观测到的掷硬币A的结果，是隐变量(不可观测)；$\theta = (\pi, p, q)$模型参数。实验一次/轮(先抛一次硬币A，然后B或者C)，得到$y$概率为：

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

【注】 对$P(y \mid z, \theta)$讨论

$z = 1$时，A正面，$P(z \mid \theta) = \pi$；第二次选硬币B。$P(y \mid z, \theta)$对$y$分情况

$y = 1$，$P(y \mid z, \theta) = p = p^y (1-p)^{1-y}$;

$y = 0$，$P(y \mid z, \theta) = 1 - p = p^y (1-p)^{1-y}$；

总之$P(y \mid z, \theta) = p^y (1-p)^{1-y}$；即， $z = 1$时， $P(z \mid \theta) P(y \mid z, \theta) = \pi p^y (1-p)^{1-y}$

同理，$z = 0$，A反面，$P(z \mid \theta) = 1 - \pi$；第二次选硬币C，$P(y \mid z, \theta) = q^y (1-q)^{1-y}$

$z = 0$时，$(z \mid \theta) P(y \mid z, \theta) = (1-\pi) q^y (1-q)^{1-y}$

# 三硬币模型

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

$$P(Y\mid\theta)=\sum_{Z}P(Z\mid\theta)P(Y\mid Z,\theta)$$

即

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

模型参数$\theta=(\pi,p,q)$的极大似然估计，即

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

此问题没有解析解，只能迭代求解。EM算法为此类问题的一种迭代算法

# EM方法 – E步

初值：$\theta^{(0)}=\left(\pi^{(0)},p^{(0)},q^{(0)}\right)$， $\mathrm{A,B,C}$正面出现的概率分别是$\pi,p,q$

第$i$步：$\theta^{(i)}=\left(\pi^{(i)},p^{(i)},q^{(i)}\right)$

EM算法第$\mathrm{i}+1$次迭代：

E步：给定参数$\theta^{(i)}$下，对观测数据$y_j$，其结果来自掷硬币B的概率$\mu_j^{(i+1)}$

【估算隐变量$Z$的状态，出现B占出现硬币B以及硬币C的总和的比例】

【E步估算观测值相关的期望，此处根据当前参数估计观测值的（平均的）状态】

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

- 【注】
  - 来自掷硬币B：$P(z\mid\theta)P(y\mid z,\theta)=\pi p^{y}(1-p)^{1-y}$；
  - 来自掷硬币C：$P(z\mid\theta)P(y\mid z,\theta)=(1-\pi)q^{y}(1-q)^{1-y}$

# EM方法 – M步

初值：$\theta^{(0)} = \left(\pi^{(0)}, p^{(0)}, q^{(0)}\right)$， $A, B, C$正面出现的概率分别是$\pi, p, q$

第$i$步：$\theta^{(i)} = \left(\pi^{(i)}, p^{(i)}, q^{(i)}\right)$

EM算法第i + 1次迭代：

M步： 估算$\theta^{(i+1)} = \left(\pi^{(i+1)}, p^{(i+1)}, q^{(i+1)}\right)$

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

- 【注】：古典概率计算
  - A正面概率$\pi^{(i+1)}$：出现正面的次数$(\sum_{j=1}^{n}\mu_j^{(i+1)})$/总次数;
  - B正面的概率：出现掷硬币B的情形$(\sum_{j=1}^{n}\mu_j^{(i+1)})$，其中结果为正面$(\sum_{j=1}^{n}\mu_j^{(i+1)}y_j)$
  - C正面的概率：出现掷硬币C的情形$(\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right))$，其中结果为正面$(\sum_{j=1}^{n}\left(1-\mu_j^{(i+1)}\right)y_j)$

# EM方法

- 记号

  观察数据(observerd data, incomplete data) $Y$

  隐藏数据(unobserved data, latent variable) $Z$

  完全数据(complete-data) $(Y, Z)$

  模型参数$\theta$

- 给定观测数据$Y$，其概率分布是$P(Y \mid \theta)$，$Y$和$Z$的联合概率分布$P(Y, Z \mid \theta)$
  - 不完全数据$Y$的似然函数$P(Y \mid \theta)$，对数似然函数$L(\theta) = \log P(Y \mid \theta)$
  - 完全数据的对数似然函数是$\log P(Y, Z \mid \theta)$
- 极大化$\log P(Y \mid \theta)$
  - 挑战：包含了有未观测数据并有包含和(或积分)的对数

# EM算法

EM算法通过迭代方式，对$L(\theta)=\log P(Y\mid\theta)$极大似然估计。

每次迭代：E步求期望；M步求极大化【单调递增，且有上界=>收敛】

**【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$P(Y,Z\mid\theta)$，条件分布$P(Z\mid Y,\theta)$

输出：模型参数$\theta$

1)选择参数的初值$\theta^{(0)}$

2)E步：记$\theta^{(i)}$为第$i$次迭代参数$\theta$的估计值，在第$i+1$次迭代的E步，计算

$$Q\left(\theta,\theta^{(i)}\right)=E_Z\left[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}\right]=\sum_Z\log P(Y,Z\mid\theta)P\left(Z\mid Y,\theta^{(i)}\right)$$

【注】基于$Z$ 的完全数据的似然函数的期望。即以$Z$的分布作为权重；$P\left(Z\mid Y,\theta^{(i)}\right)$在给定观测数据$Y$和当前的参数估计$\theta^{(i)}$下，估算出来的隐变量$Z$的条件概率分布。$Y$是由$Z$决定的

3)M步：求使$Q\left(\theta,\theta^{(i)}\right)$极大化的$\theta$。确定第$i+1$次迭代的参数的估计值$\theta^{(i+1)}$

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

4)重复第2)步和第3)步，直到收敛

# $Q$函数定义

【**定义9.1(Q函数)**】 完全数据的对数似然函数$\log P(Y,Z\mid\theta)$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$P\left(Z\mid Y,\theta^{(i)}\right)$的期望称为$Q$函数，即

$$Q\left(\theta,\theta^{(i)}\right)=E_Z\left[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}\right]$$

【说明】步骤1)参数的初值。可以任意选择，但EM算法对初值敏感

【注】 $Q\left(\theta,\theta^{(i)}\right)=E_Z\left[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}\right]=\sum_Z P\left(Z\mid Y,\theta^{(i)}\right)\log P(Y,Z\mid\theta)$

步骤2)E步求$Q\left(\theta,\theta^{(i)}\right)$

$Q$函数式中$Z$是未观测数据，$Y$是观测数据。$Q\left(\theta,\theta^{(i)}\right)$的第1个变元表示要极大化的参数，第2个变元为参数的当前估计值。每次迭代实际在求$Q$函数及其极大化。

步骤3)M步求$Q\left(\theta,\theta^{(i)}\right)$的极大化。得到$\theta^{(i+1)}$，完成一次迭代$\theta^{(i)}\to\theta^{(i+1)}$。

【注】可以证明每次迭代使似然函数增大或达到局部极值。

步骤4)给出停止迭代的条件。一般地，对较小的正数$\varepsilon_1,\varepsilon_2$，若满足

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

则停止迭代。

# EM算法的导出

通过近似求解观测数据的对数似然函数的极大化问题来导出EM算法。

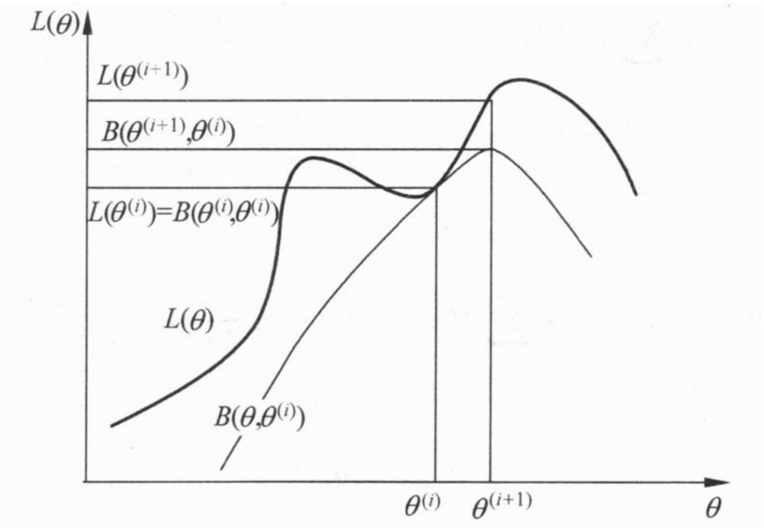
对一个含有隐变量的概率模型，目标是**极大化观测数据 (不完全数据)$Y$关于参数$\theta$的似然函数：**

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

难点：包含未观测数据的计算

EM通过迭代逐步近似极大化$L(\theta)$

$L(\theta^{(i)})$：第$i$次$\theta^{(i)}$



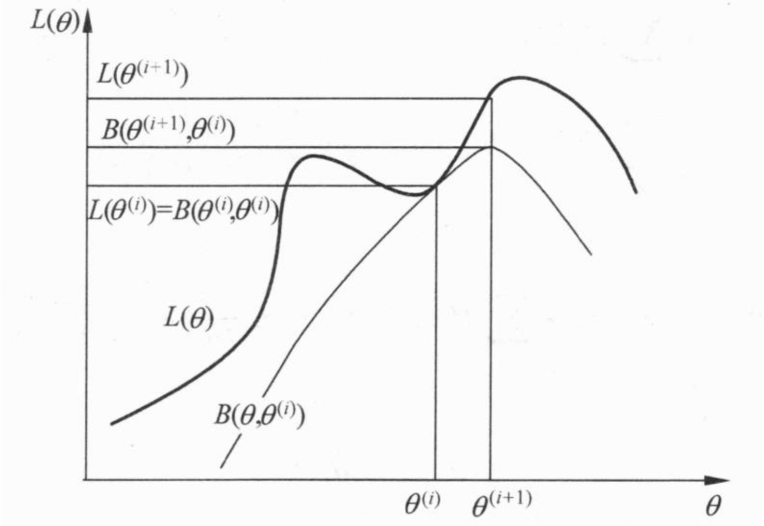
# EM算法的导出

$L(\theta)$：新估计值，期望$L(\theta)$越来越大

**迭代计算策略【单调递增，且有上界】**

从$L\left(\theta^{(\mathrm{i})}\right)$出发，估算相关的下界$B\left(\theta,\theta^{(i)}\right)$极大值

得到新的$L\left(\theta^{(i+1)}\right)$



# EM算法的导出

二者的差：$L(\theta)-L(\theta^{(i)})=\log(\sum_Z P(Y|Z,\theta)P(Z|\theta))-\log P\left(Y\middle|\theta^{(i)}\right)$

利用Jason不等式估计下界：

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

$$\geqslant\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}-\log P(Y\mid\theta^{(i)})$$

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

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

整理上式，$\mathrm{L}(\theta)\geqslant L(\theta^{(i)})+\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}$

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

则$L(\theta)\geqslant B(\theta,\theta^{(i)})$，令$\theta=\theta^{(i)}$，有$B(\theta^{(i)},\theta^{(i)})=L(\theta^{(i)})$

# EM算法的解释

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。为了使$L(\theta)$有尽可能大的增长选择

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

# EM算法的解释

现在求$\theta^{(i+1)}$的表达式。省去与$\theta$无关的项

$$\theta^{(i+1)} = \arg = \max_{\theta}\left(L(\theta^{(i)}) + \sum_Z \frac{P\left(Z|Y,\theta^{(i)}\right)\log\left(P(Y|Z,\theta)P(Z|\theta)\right)}{P\left(Z|Y,\theta^{(i)}\right)P\left(Y|\theta^{(i)}\right)}\right)$$

$$= \arg\max_{\theta}\left(\log\left(\sum_Z P\left(Y|Z,\theta^{(i)}\right)P\left(Z|\theta^{(i)}\right)\right) + \sum_Z \frac{P\left(Z|Y,\theta^{(i)}\right)\log\left(P(Y|Z,\theta)P(Z|\theta)\right)}{P\left(Z|Y,\theta^{(i)}\right)P\left(Y|\theta^{(i)}\right)}\right)$$

$$= \arg\max_{\theta}\left(\sum_Z P\left(Z \mid Y,\theta^{(i)}\right)\log\left(P(Y \mid Z,\theta)P(Z \mid \theta)\right)\right)$$

$$= \arg\max_{\theta}\left(\sum_Z P\left(Z \mid Y,\theta^{(i)}\right)\log P(Y,Z \mid \theta)\right)$$

$$= \arg\max_{\theta} Q\left(\theta,\theta^{(i)}\right)$$

等价于EM算法的一次迭代，即求$Q$函数及其极大化

# EM在无监督学习中的应用

- 监督学习是由训练数据$\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$学习条件概率分布$P(Y \mid X)$或决策函数$Y = f(X)$作为模型。训练数据中的每个样本点由输入和输出对组成

- 训练数据只有输入没有对应的输出$\{(x_1,.),(x_2,.),\ldots(x_N,.)\}$，从这样的数据学习模型称为无监督学习问题

- EM算法可以用于生成模型的无监督学习。生成模型由联合概率分布$P(X,Y)$表示，可以认为无监督学习训练数据是联合概率分布产生的数据。 $X$为观测数据， $Y$为未观测数据

# 2 EM算法的收敛性

# EM算法的收敛性

EM算法

提供一种近似计算含有隐变量概率模型的极大似然估计的方法

最大优点：简单性和普适性

问题：

EM算法得到的估计序列是否收敛？

如果收敛，是否是全局极大值或局部极大值？

# EM算法的收敛性

**定理 9.1** 设 $P(Y \mid \theta)$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，$P\left(Y \mid \theta^{(i)}\right)(i=1,2,\cdots)$为对应的似然函数序列，则$P\left(Y \mid \theta^{(i)}\right)$是单调递增

$$P\left(Y \mid \theta^{(i+1)}\right) \geqslant P\left(Y \mid \theta^{(i)}\right)$$

【注】单调性

**定理9.2** 设$L(\theta)=\log P(Y \mid \theta)$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，$L\left(\theta^{(i)}\right)(i=1,2,\cdots)$ 为对应的对数似然函数序列。

1)如果$P(Y \mid \theta)$有上界，则$L\left(\theta^{(i)}\right)=\log P\left(Y \mid \theta^{(i)}\right)$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$\theta^{*}$是$L(\theta)$的稳定点。

【注】1)单调递增且有上界，收敛

# 3 EM算法在高斯混合模型学习中的应用

# 高斯混合模型

**【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：

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

其中，$\alpha_k \geqslant 0$, $\sum_{k=1}^{K} \alpha_k = 1$， $\phi(y \mid \theta_k)$是高斯分布密度，$\theta_k = (\mu_k, \sigma_k^2)$,

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

称为第$k$个分模型。

# 完全数据的对数似然函数

假设观测数据$y_1, y_2, \cdots, y_N$由高斯混合模型生成

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

其中,$\theta = (\alpha_1, \alpha_2, \cdots, \alpha_K; \theta_1, \theta_2, \cdots, \theta_K)$。用EM算法估计高斯混合模型的参数$\theta$

【注意】混合不是多个模型加权平均，而是每次以概率选择多个模型之一，何整体效果混合

1) 明确隐变量，写出完全数据的对数似然函数

观测数据$y_j, j = 1,2,\cdots,N$

【依概率$\alpha_k$选择第$k$个高斯分布，然后依第$k$个模型的概率分布$\phi(y \mid \theta_k)$生成观测数据$y_j$】

隐变量$\gamma_{jk}$：$y_j$来自第$k$个分模型

$$\gamma_{jk} = \begin{cases} 1, & \text{第}j\text{个观测来自第}k\text{个分模型} \\ 0, & \text{否则} \end{cases} \quad j = 1,2,\cdots,N; \quad k = 1,2,\cdots,K$$

则第$j$个的完全数据为$(y_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jK})$, $j = 1,2,\cdots,N$

# 完全数据的对数似然函数

完全数据的似然函数：

【注】根据$\gamma_{jk}$定义，$P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}\mid\theta)=\prod_{k=1}^{K}[\alpha_k\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$（只有某一项$\gamma_{jk}$非0，$\alpha_k\phi(y_j\mid\theta_k)$）

$$P(y,\gamma\mid\theta)=\prod_{j=1}^{N}P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}\mid\theta)=\prod_{j=1}^{N}\prod_{k=1}^{K}[\alpha_k\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$$

$$=\prod_{k=1}^{K}\prod_{j=1}^{N}[\alpha_k\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$$

$$=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}[\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$$

$$=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}$$

其中，属于模型$k$的样本数$n_k=\sum_{j=1}^{N}\gamma_{jk}$，$\sum_{k=1}^{K}n_k=N$，那么

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

# E step

2) E step：确定$Q$函数。【注】$n_k=\sum_{j=1}^{N}\gamma_{jk}$, $E[cX]=cE[X]$, 根据概率定义$E[\gamma_{jk}]=P(\gamma_{jk}=1)$

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

$$
\hat{\gamma}_{jk}=E[\gamma_{jk}|y,\theta]=P\left(\gamma_{jk}=1\mid y,\theta\right)=\frac{P\left(\gamma_{jk}=1,y_j\mid\theta\right)}{\sum_{k=1}^{K}P\left(\gamma_{jk}=1,y_j\mid\theta\right)}=\frac{P\left(y_j\mid\gamma_{jk}=1,\theta\right)P\left(\gamma_{jk}=1\mid\theta\right)}{\sum_{k=1}^{K}P\left(y_j\mid\gamma_{jk}=1,\theta\right)P\left(\gamma_{jk}=1\mid\theta\right)\left(\gamma_{jk}\mid y,\theta\right)}
$$

$\hat{\gamma}_{jk}$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$k$对观测数据$y_j$的响应度

将$\hat{\gamma}_{jk}=E[\gamma_{jk}]$及$n_k=\sum_{j=1}^{N}E[\gamma_{jk}]$代入式$Q\left(\theta,\theta^{(i)}\right)$,得

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

# M step

3) 确定EM算法的M步

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

其中$\theta^{(i+1)} = (\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\alpha}_k,\ k = 1, \cdots, K)$ 。将$Q\left(\theta, \theta^{(i)}\right)$分别对$\mu_k, \sigma_k^2$求偏导数并令其为0，求$\hat{\alpha}_k$是在$\sum_{k=1}^{K} \alpha_k = 1$条件下求偏导并令其为0，得到

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

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

$$\hat{\alpha}_k, = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \quad k = 1,2, \cdots, K$$

重复2)-3)，直到对数似然函数值不再有明显的变化为止

# 高斯混合模型参数估计的EM算法

**【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

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

3)M步：计算新一轮迭代的模型参数

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

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

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \quad k = 1,2,\cdots,K$$

4)重复第(2)步和第(3)步，直到收敛。

# 4 EM算法的推广

# EM算法的推广

- EM 算法还可以解释为F函数(F function) 的极大极大算(maximization algorithm) , 基于这个解释有若干变形与推广，如广义期望极大(generalized expectation maximization, GEM) 算法