吉林大学

统计计算

Statistical Computation

机器学习研究室
统计计算课程组

第4章统计计算中的优化问题

- 4.1 最大似然估计
- 4.2 非线性回归
- 4.3 EM算法

4.1 最大似然估计

4.1.1 概述
4.1.2 得分法
4.1.3 精简最大似然估计

4.1.1 概述

设总体$\mathbf{X}$有密度或概率函数$p(\mathbf{x}|\boldsymbol{\theta})$,

$\boldsymbol{\theta}$为m维的分布参数。

一组样本$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$,

则**似然函数**（likelihood function）为

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(\mathbf{X}_i|\boldsymbol{\theta}\right),$$

4.1.1 概述

- 对数似然函数为

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{X}_i|\boldsymbol{\theta}),$$

4.1.1 概述

- 最大似然估计求解过程：

（1） 写出似然函数；

（2） 对似然函数取对数，并整理；

（3） 求导数；

（4） 解似然方程

4.1.2 得分法

设对数似然函数$l(\boldsymbol{\theta})$有二阶连续偏导数

- $\nabla l(\boldsymbol{\theta})$称为**得分函数**(score function)
- 方程$\nabla l(\boldsymbol{\theta}) = \mathbf{0}$，称为**估计方程**
- 求最大似然估计可以通过解估计方程

4.1.2 得分法

设**参数真值为**$\boldsymbol{\theta}_\star$,

$l(\boldsymbol{\theta})$**最大值点为**$\widehat{\boldsymbol{\theta}}$,

- **X的信息阵**

$$I(\boldsymbol{\theta}_\star) = \mathrm{Var}\left(\nabla \ln p(X|\boldsymbol{\theta}_\star)\right) = \frac{1}{n}\mathrm{Var}\left(\nabla l(\boldsymbol{\theta}_\star)\right) = E[-\nabla^2 \ln p(X|\boldsymbol{\theta}_\star)] = \frac{1}{n}E[-\nabla^2 l(\boldsymbol{\theta}_\star)]$$

- 在适当正则性条件下 $\widehat{\boldsymbol{\theta}}$渐近正态分布:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_\star)\right),\ n \to \infty,$$

4.1.2 得分法

记$I_n(\boldsymbol{\theta}_\star) = \mathrm{Var}\left(\nabla l(\boldsymbol{\theta}_\star)\right)$

- 称$I_n(\boldsymbol{\theta}_\star)$为$(X_1, X_2, \ldots, X_n)$的**信息阵**
- 当$X_1, X_2, \ldots, X_n$独立同分布时$I_n(\boldsymbol{\theta}_\star) = nI(\boldsymbol{\theta}_\star)$
- 求得最大似然估计$\hat{\boldsymbol{\theta}}$后，用$[I_n\left(\hat{\boldsymbol{\theta}}\right)]^{-1}$估计$\hat{\boldsymbol{\theta}}$协方差阵

4.1.2 得分法

$l(\boldsymbol{\theta})$最大值点$\hat{\boldsymbol{\theta}}$的数值求解：牛顿法、BFGS法等

- **牛顿法**： $\hat{\boldsymbol{\theta}}$的协方差阵用$\left[-\nabla^2 l(\hat{\boldsymbol{\theta}})\right]^{-1}$来估计
- 当样本量充分大且$\boldsymbol{\theta}$接近于真值$\boldsymbol{\theta}_*$时 海色阵$\nabla^2 l(\boldsymbol{\theta}) \approx -I_n(\boldsymbol{\theta})$，
- **牛顿法的迭代公式**为

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)}\left[-I_n\left(\hat{\boldsymbol{\theta}}^{(t)}\right)\right]^{-1}\nabla l\left(\hat{\boldsymbol{\theta}}^{(t)}\right).$$

用负信息阵代替对数似然函数的海色阵

这种方法叫做**得分法**(scoring)。

4.1.2 得分法

线性回归模型：

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中：$\boldsymbol{\varepsilon}|\mathrm{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ ，

被解释变量的条件分布为 $\boldsymbol{y}|\mathrm{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$

其条件概率密度函数为：

$$f(\boldsymbol{y}|\mathbf{X}) = \left(2\pi\sigma^2\right)^{-\mathrm{n}/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

用 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\sigma}^2$ 代入，取对数，于是LF为：

$$\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

极大似然估计就是要求 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\sigma}^2$ 使得 $\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ 最大。

令$e = (Y - X\beta)$

$$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} \qquad \hat{\sigma}^2_{ML} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n} \neq \hat{\sigma}^2_{OLS} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K} = s^2$$

4.1.2 得分法

例39.1 (逻辑斯谛回归参数估计)

设$Y_i \sim \mathrm{B}(m_i, \pi_i), i = 1,2,\ldots,n, Y_1, Y_2, \ldots, Y_n$相互独立，

其中：

$\pi_i = \exp(\boldsymbol{\beta}^T \mathbf{x}_i)/[1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)]$, $\boldsymbol{\beta}$为未知参数向量，

$\mathbf{x}_i, i = 1,2,\ldots,n$是已知的自变量

称为**逻辑斯谛回归模型**

函数$\mathrm{logit}(\pi) \triangleq \log\frac{\pi}{1-\pi}, \pi \in (0,1)$称为**逻辑斯谛函数**

4.1.2 得分法

- Y_i的概率函数

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

$$= \binom{m_i}{y_i} [1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)]^{-m_i} [\exp(\boldsymbol{\beta}^T \mathbf{x}_i)]^{y_i},$$

,

- 其对数为

$$\log P(Y_i = y_i) = \log \binom{m_i}{y_i} + y_i \cdot \boldsymbol{\beta}^T \mathbf{x}_i - m_i \log[1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)]$$

4.1.2 得分法

- **对数似然函数**（省略了不随$\boldsymbol{\beta}$变化的部分）

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \cdot \boldsymbol{\beta}^T \mathbf{x}_i - m_i \log[1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)]\}$$

- **梯度**

$$\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i - \frac{m_i \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \mathbf{x}_i \right) = \sum_{i=1}^{n} (y_i - m_i \pi_i) \mathbf{x}_i,$$

- **海色阵**

$$\nabla^2 l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} m_i \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)]^2} \cdot \mathbf{x}_i \mathbf{x}_i^T$$

$$= -\sum_{i=1}^{n} m_i \, \pi_i (1 - \pi_i) \mathbf{x}_i \mathbf{x}_i^T .$$

4.1.2 得分法

- 得分法的迭代公式

$$
\begin{cases}
\pi_i^{(t)} = \text{logit}^{-1}\left(\left[\boldsymbol{\beta}^{(t)}\right]^T \mathbf{x}_i\right),\ i = 1,2,\dots,n, \\
\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[-\sum_{i=1}^{n} m_i\, \pi_i^{(t)}\left(1-\pi_i^{(t)}\right)\mathbf{x}_i\mathbf{x}_i^T\right]^{-1}\left[\sum_{i=1}^{n}\left(y_i - m_i\pi_i^{(t)}\right)\mathbf{x}_i\right], \\
t = 0,1,2,\dots
\end{cases}
$$

4.1.3 精简最大似然估计

设参数$\boldsymbol{\theta}$分为$\boldsymbol{\theta}_1$和$\boldsymbol{\theta}_2$两部分,

如果对任意$\boldsymbol{\theta}_1$， 都能比较容易地求得

$$\underset{\boldsymbol{\theta}_2}{\operatorname{argmax}} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1),$$

则 $$\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1} \max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1))$$

问题简化为以$\boldsymbol{\theta}_1$为自变量的优化问题。

这样的方法称为**精简最大似然方法(concentrated MLE)**。

4.2 非线性回归

4.3 EM算法

4.3.1 EM算法引入
4.3.2 EM算法的收敛性
4.3.3 EM算法的应用

4.3.1 EM算法引入

- EM算法包括
  （1）E步（Expectation期望步）
  （2）M步（Maximization最大化步）

- EM算法主要用来求解
  （1）存在隐变量
  （2）数据缺失

的极大似然估计问题。

- Dempster, Laird , Rubin “Maximum Likelihood from Incomplete Data via the EM Algorithm”

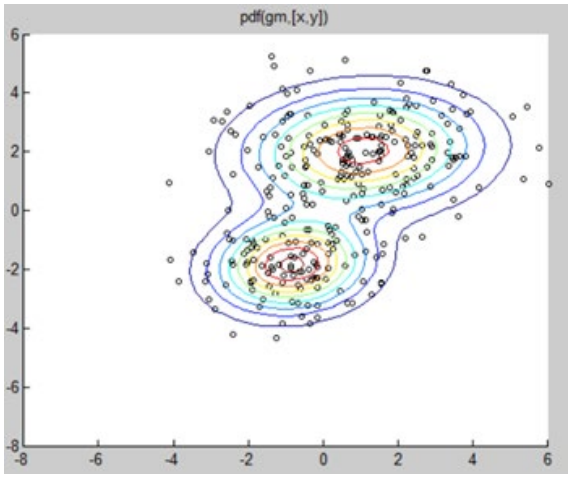
4.3.1 EM算法引入

- 高斯混合模型（Gaussian mixture model，简称GMM）
- 隐式马尔科夫（HMM）算法
- LDA主题模型的变分推断
- VAE算法推导的前置知识
- K-MEANS

4.3.1 EM算法引入

4.3.1 EM算法引入

- 100个男、女身高，分布？男多少？女多少？
- 采用混合高斯模型，假设男和女的分布都是符合高斯分布的，然后给定这个高斯分布一个初始值，这样这个高斯分布就是已知的。
- 用这个已知的高斯分布来估计男的多少人，女的多少人，假设男和女的类别分布为Q(z)，可以求Q(z)的期望，用期望来表示下一次迭代类别的初始值，就知道男和女的所属类别，可用最大似然函数来估新的高斯模型的参数，重复上述步骤…直到收敛！



4.3.1 EM算法引入

双因素模型：因素A和因素B

	B_1	B_2	B_3
A_1	10	15	17
A_2	22	23	NA

估计缺失数据。

4.3.1 EM算法引入

设双因素模型为$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$，其中：

- x_{ij}表示(A因素的i；B因素的j)数值；
- μ 表示常数项；
- α_i表示A因素的项；β_j表示B因素的项；
- 约束：$\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 + \beta_3 = 0$
- ϵ_{ij}表示扰动项；$\epsilon_{ij} \sim N(0, \sigma^2)$

则$x_{ij} \sim N\left(\mu + \alpha_i + \beta_j, \sigma^2\right)$

要估计缺失数据为$\hat{x}_{23}$

解法1：MLE

为此，使用MLE先估计 $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j$

似然函数为：

$$L(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{11}-\mu-\alpha_1-\beta_1)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{12}-\mu-\alpha_1-\beta_2)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{13}-\mu-\alpha_1-\beta_3)^2}{2\sigma^2}}$$
$$\times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{21}-\mu-\alpha_2-\beta_1)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{22}-\mu-\alpha_2-\beta_2)^2}{2\sigma^2}}$$

对数似然函数要估计的部分：

$$Q(\theta) = (x_{11} - \mu - \alpha_1 - \beta_1)^2 + (x_{12} - \mu - \alpha_1 - \beta_2)^2 + (x_{13} - \mu - \alpha_1 - \beta_3)^2 + (x_{21} - \mu - \alpha_2 - \beta_1)^2 + (x_{22} - \mu - \alpha_2 - \beta_2)^2$$

在约束：$\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 + \beta_3 = 0$下

求关于θ的梯度：

$$\begin{aligned}\frac{\partial Q}{\partial \mu} =& -2(x_{11} - \mu - \alpha_1 - \beta_1) - 2(x_{12} - \mu - \alpha_1 - \beta_2) - 2(x_{13} - \mu - \alpha_1 - \beta_3) \\ & - 2(x_{21} - \mu - \alpha_2 - \beta_1) - 2(x_{22} - \mu - \alpha_2 - \beta_2) \\ =& -2(x_{11} + x_{12} + x_{13} + x_{21} + x_{22} - 5\mu - 3\alpha_1 - 2\alpha_2 - 2\beta_1 - 2\beta_2 - \beta_3) \\ =& 0\end{aligned}$$

则$87 - 5\mu - \alpha_1 + \beta_3 = 0$　　（1）

$$\frac{\partial Q}{\partial \alpha_1} = -2(x_{11} - \mu - \alpha_1 - \beta_1) - 2(x_{12} - \mu - \alpha_1 - \beta_2) - 2(x_{13} - \mu - \alpha_1 - \beta_3)$$
$$= -2(x_{11} + x_{12} + x_{13} - 3\mu - 3\alpha_1 - \beta_1 - \beta_2 - \beta_3)$$
$$= 0$$

则$14 - \mu - \alpha_1 = 0$ （2）

$$\frac{\partial Q}{\partial \alpha_2} = -2(x_{21} - \mu - \alpha_2 - \beta_1) - 2(x_{22} - \mu - \alpha_2 - \beta_2) = 0$$

则$45 - 2\mu - 2\alpha_2 + \beta_3 = 0$ (3)

$$\frac{\partial Q}{\partial \beta_1} = -2(x_{11} - \mu - \alpha_1 - \beta_1) - 2(x_{21} - \mu - \alpha_2 - \beta_1) = 0$$

则$16 - \mu - \beta_1 = 0$ (4)

$$\frac{\partial Q}{\partial \beta_2} = -2(x_{12} - \mu - \alpha_1 - \beta_2) - 2(x_{22} - \mu - \alpha_2 - \beta_2) = 0$$

则$19 - \mu - \beta_2 = 0$ (5)

$$\frac{\partial Q}{\partial \beta_3} = -2(x_{13} - \mu - \alpha_1 - \beta_3)$$

则$17 - \mu - \alpha_1 - \beta_3 = 0$ (6)

综合以上，得θ的估计值为：

$$\hat{\mu}=19,\hat{\alpha}_1=-5,\hat{\alpha}_2=5,\hat{\beta}_1=-3,\hat{\beta}_2=0,\hat{\beta}_3=3$$

估计缺失数据为$\hat{x}_{23}=\hat{\mu}+\hat{\alpha}_2+\hat{\beta}_3=27$

解法2：EM

完全数据时：

对数似然函数要估计的部分：

$$Q(\theta) = \sum (x_{ij} - \mu - \alpha_i - \beta_j)^2$$

在约束：$\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 + \beta_3 = 0$下

求关于θ的梯度：

$$\frac{\partial Q}{\partial \mu} = -2 \sum (x_{ij} - \mu - \alpha_i - \beta_j) = 0$$

则$\sum x_{ij} - 6\mu = 0$ (W1)

解法2：EM

$$\frac{\partial Q}{\partial \alpha_1} = -2(x_{11} - \mu - \alpha_1 - \beta_1) - 2(x_{12} - \mu - \alpha_1 - \beta_2) - 2(x_{13} - \mu - \alpha_1 - \beta_3)$$
$$= -2(x_{11} + x_{12} + x_{13} - 3\mu - 3\alpha_1 - \beta_1 - \beta_2 - \beta_3) = 0$$

则 $\frac{x_{11}+x_{12}+x_{13}}{3} - \mu - \alpha_1 = 0$ (W2)

$$\frac{\partial Q}{\partial \alpha_2} = -2(x_{21} - \mu - \alpha_2 - \beta_1) - 2(x_{22} - \mu - \alpha_2 - \beta_2) - 2(x_{23} - \mu - \alpha_2 - \beta_3) = 0$$

则 $\frac{x_{21}+x_{22}+x_{22}}{3} - \mu - \alpha_2 = 0$ (W3)

$$\frac{\partial Q}{\partial \beta_1} = -2(x_{11} - \mu - \alpha_1 - \beta_1) - 2(x_{21} - \mu - \alpha_2 - \beta_1) = 0$$

则 $\frac{x_{11}+x_{21}}{2} - \mu - \beta_1 = 0$ (W4)

解法2：EM

$$\frac{\partial Q}{\partial \beta_2} = -2(x_{12} - \mu - \alpha_1 - \beta_2) - 2(x_{22} - \mu - \alpha_2 - \beta_2) = 0$$

则 $\frac{x_{12}+x_{22}}{2} - \mu - \beta_2 = 0$ (W5)

$$\frac{\partial Q}{\partial \beta_3} = -2(x_{13} - \mu - \alpha_1 - \beta_3) - 2(x_{23} - \mu - \alpha_2 - \beta_3)$$

则 $\frac{x_{13}+x_{23}}{2} - \mu - \beta_3 = 0$ (W6)

于是

$$\hat{\mu} = \bar{x}, \hat{\alpha}_i = \bar{x}_{i.} - \bar{x}, \hat{\beta}_j = \bar{x}_{.j} - \bar{x}$$

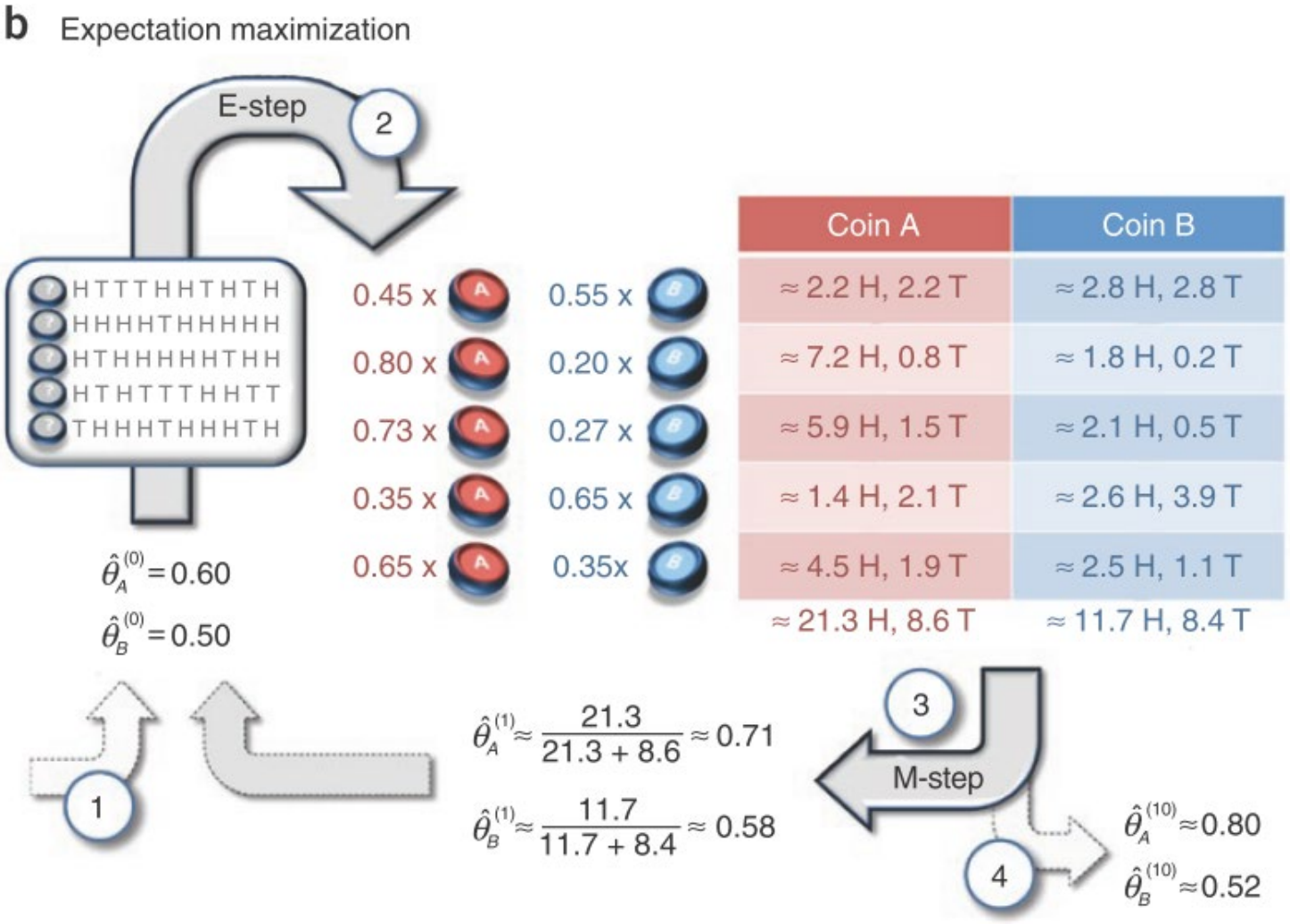
解法2：EM

1，给定初始的$\hat{x}_{23}$
2，用完全数据的公式估计新的$\hat{x}_{23}$
3，与初始的比较，接近则停止，否则迭代

4.3.1 EM算法引入

Chuong B Do, Serafim Batzoglou. What is the expectation maximization algorithm? Nature Biotechnology, 26, 897-899 (2008).

4.3.1 EM算法引入


b Expectation maximization
E-step
2
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
0.45 x
0.55 x
0.80 x
0.20 x
0.73 x
0.27 x
0.35 x
0.65 x
0.65 x
0.35x
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

算法9.1 EM算法

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

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] \\ &= \sum_Z \log P(Y,Z|\theta) P(Z|Y,\theta^{(i)}) \end{aligned} \tag{9.9}$$

这里，$P(Z|Y,\theta^{(i)})$ 是在给定观测数据 Y 和当前的参数估计 $\theta^{(i)}$ 下隐变量数据 Z 的条件概率分布；

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

$$\theta^{(i+1)} = \arg\max_\theta Q(\theta,\theta^{(i)}) \tag{9.10}$$

（4）重复第（2）步和第（3）步，直到收敛

EM算法

- Q函数定义：

完全数据的对数似然函数logP(Y,Z|Θ)关于在给定观测数据Y和当前函数$\Theta^{(i)}$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：

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

EM算法

- 算法说明：
- 步骤3，完成一次迭代：$\Theta^{(i)}$到$\Theta^{(i+1)}$，将证明每次迭代使似然函数增大或达到局部最大值。
- 步骤4，停止迭代的条件

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

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、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」**

设随机变量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。

则三硬币模型

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

三硬币模型

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

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

- 即：$$P(Y\mid\theta)=\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}]$$
- 极大似然估计：

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

- 该问题没有解析解，使用EM迭代法：

EM方法

- 选取初值： $\theta^{(0)} = (\pi^{(0)}, p^{(0)}, q^{(0)})$
- 第i步的估计值： $\theta^{(i)} = (\pi^{(i)}, p^{(i)}, q^{(i)})$
- EM算法第i+1次迭代：
- 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：

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

- M步: 计算模型参数的新估计值

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

EM方法

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

对 $y_j=1$ 与 $y_j=0$ 均有 $\mu_j^{(1)}=0.5$

利用迭代公式，得：

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

$\mu_j^{(2)}=0.5$， $j=1,2,\cdots,10$

继续迭代，得：

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

得到模型参数的极大似然估计：

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

EM方法

如果取初值：

$$\pi^{(0)} = 0.4, \quad p^{(0)} = 0.6, \quad q^{(0)} = 0.7$$

$$\hat{\pi} = 0.4064, \quad \hat{p} = 0.5368, \quad \hat{q} = 0.6432$$

完全数据 complete-data $P(Y, Z \mid \theta)$

不完全数据 incomplete-data $P(Y \mid \theta)$

4.3.2 EM算法的收敛性

- 为什么EM算法能近似实现对观测数据的极大似然估计？
- 极大化(不完全数据)Y关于参数Θ的极大似然函数：

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

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

- 难点：有未观测数据，包含和的对数。
- EM通过迭代逐步近似极大化L(Θ),希望 $L(\theta) > L(\theta^{(i)})$

EM算法的导出

- 考虑二者的差：

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

- Jason不等式：

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

EM算法的导出

- 令：

$$B(\theta,\theta^{(i)}) \triangleq L(\theta^{(i)}) + \sum_{Z} P(Z \mid Y,\theta^{(i)}) \log \frac{P(Y \mid Z,\theta)P(Z \mid \theta)}{P(Z \mid Y,\theta^{(i)})P(Y \mid \theta^{(i)})}$$

- 则：

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

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

- 选择：任何可以使 $B(\theta,\theta^{(i)})$ 增大的 θ，也可以使 $L(\theta)$ 增大

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

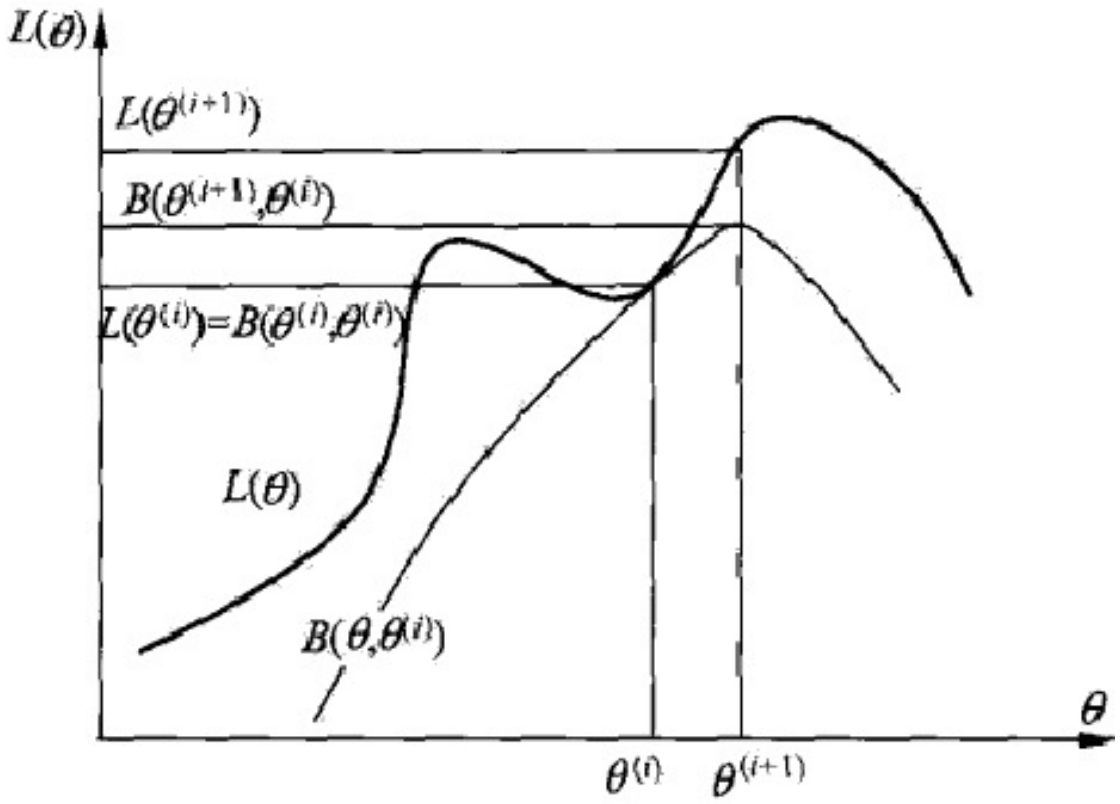
EM算法的导出

- 省去和Θ无关的项：

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

EM算法的解释

L(Θ)开始



EM算法的收敛性

- EM，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，
- EM，最大优点：简单性和普适性；
- 疑问：
- 1、EM算法得到的估计序列是否收敛？
- 2、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

- 两个收敛定理：
- **定理9.1**：设P(Y|Θ)为观测数据的似然函数，$\Theta^{(i)}(i=1,2..)$为EM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

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

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

- 由：

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

EM算法的收敛性

- 令：

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

- 则：

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

- 得：

$$\begin{aligned}&\log P(Y\mid\theta^{(i+1)})-\log P(Y\mid\theta^{(i)})\\&=[Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})]-[H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})]\end{aligned}$$

- 只需证右端非负

EM算法的收敛性

- 前半部分，$\Theta^{(i+1)}$为极大值，所以

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

- 后半部分：

$$\begin{aligned}
&H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})\\
&\quad=\sum_{Z}\left(\log\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}\right)P(Z\mid Y,\theta^{(i)})\\
&\quad\leqslant\log\left(\sum_{Z}\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}P(Z\mid Y,\theta^{(i)})\right)\\
&\quad=\log P(Z\mid Y,\theta^{(i+1)})=0
\end{aligned}$$

EM算法的收敛性

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

4.3.3 EM算法的应用

- EM算法在高斯混合模型学习中的应用

  - 高斯混合模型：
  - 概率分布模型；$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$
  - 系数：$\alpha_k\geqslant 0,\ \sum_{k=1}^{K}\alpha_k=1$
  - 高斯分布密度：$\phi(y|\theta_k)\qquad \theta_k=(\mu_k,\sigma_k^2)$
  - 第K个分模型：$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$

可任意高斯模型

高斯混合模型参数估计的EM算法

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

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

$$\theta = (\alpha_1, \alpha_2, \cdots, \alpha_K; \theta_1, \theta_2, \cdots, \theta_K)$$

- 用EM算法估计参数；
- 1、明确隐变量，写出完全数据的对数似然函数：
  - 设想观测数据yi是依概率a_k选择第k个高斯分模型 $\phi(y|\theta_k)$ 生成，隐变量

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

EM算法在高斯混合模型学习中的应用

- 1、明确隐变量，写出完全数据的对数似然函数：
- 完全数据：$(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK})$，$j=1,2,\cdots,N$
- 似然函数：

$$P(y,\gamma|\theta)=\prod_{j=1}^{N}P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}|\theta)$$

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

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

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

$$n_k=\sum_{j=1}^{N}\gamma_{jk}$$

$$\sum_{k=1}^{K}n_k=N$$

EM算法在高斯混合模型学习中的应用

- 1、明确隐变量，写出完全数据的对数似然函数：

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

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

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

需要计算 $E(\gamma_{jk}\,|\,y,\theta)$，记为 $\hat{\gamma}_{jk}$

- 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。

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

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

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

$$\text{将 } \hat{\gamma}_{jk} = E\gamma_{jk} \text{ 及 } n_k = \sum_{j=1}^{N} E\gamma_{jk} \text{ 代入}$$

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

EM算法在高斯混合模型学习中的应用

- 3、确定EM算法的M步：
- 求：

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

用 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$，$k=1,2,\cdots,K$，表示 $\theta^{(i+1)}$

- 采用求导的方法：

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

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、设定初始值开始迭代
- 2、E步，响应度计算

$$\hat{\gamma}_{jk}=\frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)}$$

高斯混合模型参数估计的EM算法

- 输入：观测数据$y_1,y_2,\ldots y_N$, 高斯混合模型
- 输出：高斯混合模型参数
- 3、M步，计算新一轮迭代的模型参数：

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

- 4、重复2，3步直到收敛