EM应用例子-GMM参数估计

EM算法（期望最大算法）

- 设一个完整数据集 $\boldsymbol{x}=\left[x_1, x_2, \cdots, x_M\right]^T$
- 存在完整数据集到观测数据（不完整集）的映射 $\boldsymbol{y}=g(\boldsymbol{x})$
- 以完整数据集的条件数学期望代替对数似然函数

$$E_{x|y}\left[\ln p_x(\boldsymbol{x}|\theta)\right]$$

$$=\int \ln p_x(\boldsymbol{x}|\theta)\, p(\boldsymbol{x}|\boldsymbol{y},\theta)\, d\boldsymbol{x}$$

- 为使积分进行，待估计参数使用猜测值，构成迭代 $\theta^{(0)} \rightarrow \theta^{(1)}, \cdots, \rightarrow \theta^{(\mathrm{k})}$

EM算法描述

第 1 步：初始化，选择$\boldsymbol{\theta}$的初始猜测值，令$m=0$，给出$\boldsymbol{\theta}^{(m)}$；

第 2 步：由观测数据$\boldsymbol{y}$和$\boldsymbol{\theta}$的猜测值$\boldsymbol{\theta}^{(m)}$，

得到完整数据集$\boldsymbol{x}$的条件概率$p\left(\boldsymbol{x} \middle| \boldsymbol{y}, \boldsymbol{\theta}^{(m)}\right)$；

第 3 步：计算完整数据集下对数似然函数$l(\theta|\boldsymbol{x}) = \log p_x(\boldsymbol{x}|\theta)$的条件期望

$$\boldsymbol{Q}\left(\boldsymbol{\theta} \middle| \boldsymbol{\theta}^{(m)}\right) = E_{x|y,\theta^{(m)}}\left[\log p_x(\boldsymbol{x}|\theta)\right] = \int \log p_x(\boldsymbol{x}|\theta)\, p(\boldsymbol{x}|\boldsymbol{y}, \theta^{(m)})\, d\boldsymbol{x}$$

续

第 4 步：求$\theta=\theta^{(m+1)}$使得$Q\left(\theta\middle|\theta^{(m)}\right)$最大，即

$$\theta^{(m+1)}=\underset{\theta\in\Omega}{\arg\max}\left\{Q\left(\theta\middle|\theta^{(m)}\right)\right\}$$

第 5 步：若满足停止条件，则$\hat{\theta}=\theta^{(m+1)}$为所得 MLE，若不满足停止条件，

令$m:=m+1$返回第 2 步。

特例1：数据缺失或隐变量情况

观测数据集 $\boldsymbol{y}$

完整数据集 $\boldsymbol{x}=\{\boldsymbol{y},\boldsymbol{z}\}$

隐藏数据集 $\boldsymbol{z}$

则有：

$$\boldsymbol{Q}\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(m)}\right)=E_{z|y,\theta^{(m)}}\left[\log p_x(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})\right]$$

$$=\int_{\mathfrak{I}}\log p_x(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})\,p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(m)})d\boldsymbol{z}$$

样本独立同分布情况

设 $p(\boldsymbol{x}|\boldsymbol{\theta})=\prod_{i=0}^{N-1} p(\boldsymbol{x}(n_i)|\boldsymbol{\theta})$,

并且 $\boldsymbol{y}(n_i)$ 仅与 $\boldsymbol{x}(n_i)$ 有关

则有
$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})=\sum_{i=0}^{N-1} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$$

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})=E_{\boldsymbol{x}(n_i)|\boldsymbol{y}(n_i),\theta^{(m)}}\left[\log p_x(\boldsymbol{x}(n_i)|\boldsymbol{\theta})\right]$$

这里
$$=\int \log p_x(\boldsymbol{x}(n_i)|\boldsymbol{\theta})\,p(\boldsymbol{x}(n_i)|\boldsymbol{y}(n_i),\boldsymbol{\theta}^{(m)})\,d\boldsymbol{x}$$

EM算法解高斯混合模型

观测向量集表示为 $\boldsymbol{y} = \left\{ \boldsymbol{y}_i \middle| 0 \le i < N \right\}$ 是i.i.d.的

GMM模型

$$p(\boldsymbol{y}_i) = \sum_{k=1}^{K} w_k N\left(\boldsymbol{y}_i \middle| \boldsymbol{\mu}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}\right)$$

约束条件

$$\sum_{k=1}^{K} w_k = 1 \qquad 0 \le w_k \le 1$$

由i.i.d.样本集估计参数集

$$\boldsymbol{\theta} = \left\{ w_k, \boldsymbol{\mu}_k, \mathbf{C}_k \middle| k = 1,2,\cdots,K \right\}$$

定义隐变量 $z_i \in \{1,2,\cdots,K\}$

隐变量向量 $\mathbf{z} = [z_0, z_1, \cdots, z_{N-1}]^T$

完整数据集 $\boldsymbol{x} = \{\boldsymbol{y}, \mathbf{z}\}$

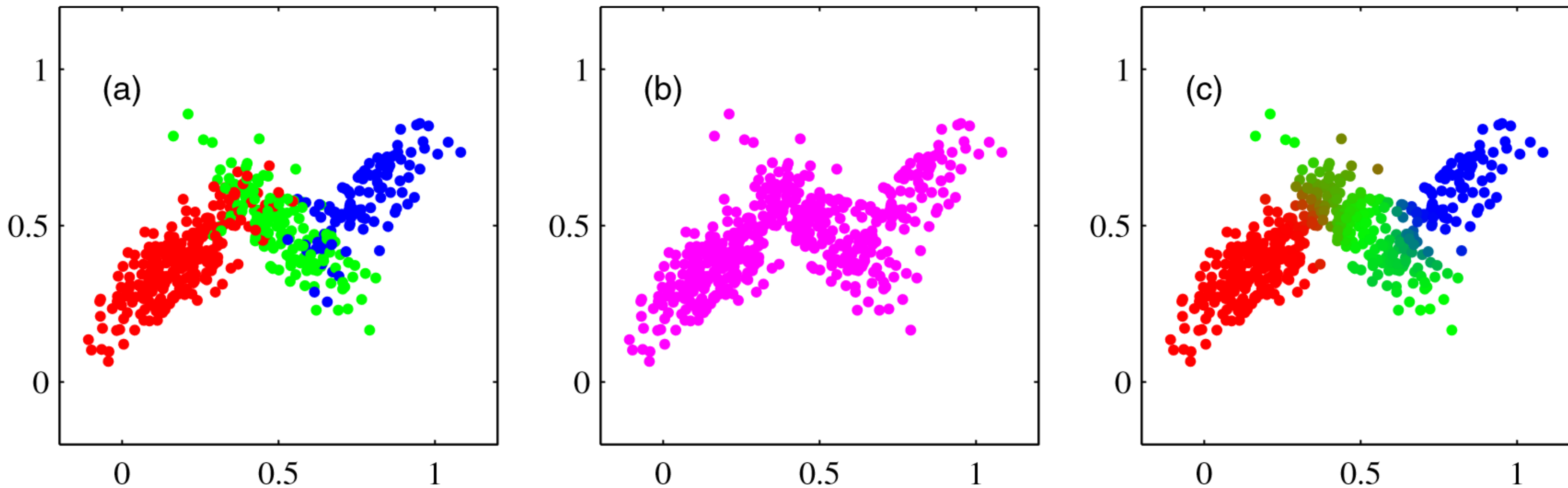
由
$$p_x(\boldsymbol{y}_i, z_i = k|\boldsymbol{\theta}) = w_k N(\boldsymbol{y}_i|\boldsymbol{\mu}_\mathbf{k}, \mathbf{C}_\mathbf{k})$$

设已得到参数向量的猜测值 $\boldsymbol{\theta}^{(m)}$ 则

$$p\left(z_i = k\middle|\boldsymbol{y}_i, \boldsymbol{\theta}^{(m)}\right) = \frac{p_x\left(\boldsymbol{y}_i, z_i = k\middle|\boldsymbol{\theta}^{(m)}\right)}{p\left(\boldsymbol{y}_i\middle|\boldsymbol{\theta}^{(m)}\right)}$$

$$= \frac{w_k^{(m)} N\left(\boldsymbol{y}_i\middle|\boldsymbol{\mu}_k^{(m)}, \mathbf{C}_k^{(m)}\right)}{\sum_{l=1}^{K} w_l^{(m)} N\left(\boldsymbol{y}_i\middle|\boldsymbol{\mu}_l^{(m)}, \mathbf{C}_l^{(m)}\right)}$$

GMM：隐变量观点示例



左：已知 π_k 和参数 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ，仿真产生若干数据，

并记下 $z_i = k$ ，用红绿蓝表示样本点产生那个k分量，

故：左图是联合分布

中：去掉颜色，即去掉隐变量信息，实际样本是不知隐变量

右：已知 π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ，用中图的各样本点坐标估计：

$$\gamma_{ik} = p\left(z_i = k \middle| \mathbf{y}_i\right)$$

<u>实际中，只有中图的样本点集：估计</u> π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

简记 $$\gamma_{ik}^{(m)} = p\left(z_i = k \middle| \boldsymbol{y}_i, \boldsymbol{\theta}^{(m)}\right)$$

$$\sum_{k=1}^{K} \gamma_{ik}^{(m)} = 1$$

由数据缺失和**i.i.d.**情况，得

$$\boldsymbol{Q_i}\left(\boldsymbol{\theta} \middle| \boldsymbol{\theta}^{(m)}\right) = E_{z_i \left| y_i, \theta^{(m)} \right.}\left[\log p_x(\boldsymbol{y_i}, z_i | \boldsymbol{\theta})\right]$$

$$= \sum_{k=1}^{K} p\left(z_i = k \middle| \boldsymbol{y}_i, \boldsymbol{\theta}^{(m)}\right) \log p_x(\boldsymbol{y_i}, z_i = k | \boldsymbol{\theta})$$

$$= \sum_{k=1}^{K} \gamma_{ik}^{(m)} \log\left(w_k N\left(\boldsymbol{y}_i \middle| \boldsymbol{\mu_k}, \mathbf{C_k}\right)\right)$$

$$= \sum_{k=1}^{K} \gamma_{ik}^{(m)} \left[\log w_k - \frac{1}{2}\log\left|\mathbf{C_k}\right| - \frac{1}{2}\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right)^T \mathbf{C_k}^{-1}\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right)\right] + C$$

E步

令 $$n_k^{(m)} = \sum_{i=0}^{N-1} \gamma_{ik}^{(m)}$$

$$\boldsymbol{Q}\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(m)}\right) = \sum_{i=0}^{N-1} \boldsymbol{Q_i}\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(m)}\right) =$$

$$\sum_{i=0}^{N-1}\sum_{k=1}^{K} \gamma_{ik}^{(m)} \left[\log w_k - \frac{1}{2}\log\left|\mathbf{C_k}\right| - \frac{1}{2}\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right)^T \mathbf{C_k}^{-1}\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right) \right]$$

$$= \sum_{k=1}^{K} n_k^{(m)} \log w_k - \frac{1}{2}\sum_{k=1}^{K} n_k^{(m)} \log\left|\mathbf{C_k}\right|$$

$$-\frac{1}{2}\sum_{i=0}^{N-1}\sum_{k=1}^{K} \gamma_{ik}^{(m)} \left[\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right)^T \mathbf{C_k}^{-1}\left(\boldsymbol{y_i} - \boldsymbol{\mu_k}\right) \right]$$

M步

分别对各参数求导，并令导数为0，得

$$w_k^{(m+1)} = \frac{n_k^{(m)}}{\sum_{l=1}^{K} n_l^{(m)}} = \frac{n_k^{(m)}}{N}$$

$$\boldsymbol{\mu}_k^{(m+1)} = \frac{1}{n_k^{(m)}} \sum_{i=0}^{N-1} \gamma_{ik}^{(m)} \mathbf{y_i}$$

$$\mathbf{C}_{\mathbf{k}}^{(m+1)} = \frac{1}{n_k^{(m)}} \sum_{i=0}^{N-1} \gamma_{ik}^{(m)} \left(\mathbf{y_i} - \boldsymbol{\mu}_{\mathbf{k}}\right)\left(\mathbf{y_i} - \boldsymbol{\mu}_{\mathbf{k}}\right)^T$$

GMM参数的EM算法完整描述

S1 初始化 $m=0$，设置停止门限 ε，

给出初值 $w_k^{(0)}$，$\boldsymbol{\mu}_k^{(0)}$，$\mathbf{C}_\mathbf{k}^{(0)}$，$k=1,2,\cdots,K$，

并计算初始对数似然函数值

$$l(\boldsymbol{\theta}^{(m)}|\mathbf{y})=\frac{1}{N}\sum_{i=0}^{N-1}\log\left(\sum_{k=1}^{K}w_k^{(m)}N\left(\mathbf{y}_i\middle|\boldsymbol{\mu}_k^{(m)},\mathbf{C}_\mathbf{k}^{(m)}\right)\right)$$

S2 E-步，对 $k=1,2,\cdots,K$ 计算

$$\gamma_{ik}^{(m)}=\frac{w_k^{(m)}N\left(\boldsymbol{y}_i\,\middle|\,\boldsymbol{\mu}_k^{(m)},\mathbf{C}_k^{(m)}\right)}{\sum_{l=1}^{K}w_l^{(m)}N\left(\boldsymbol{y}_i\,\middle|\,\boldsymbol{\mu}_l^{(m)},\mathbf{C}_l^{(m)}\right)},\quad i=0,1,\cdots,N-1$$

$$n_k^{(m)}=\sum_{i=0}^{N-1}\gamma_{ik}^{(m)}$$

S3 M 步，对 $k=1,2,\cdots,K$ 计算

$$w_k^{(m+1)}=\frac{n_k^{(m)}}{\sum_{l=1}^{K}n_l^{(m)}}=\frac{n_k^{(m)}}{N}$$

$$\boldsymbol{\mu}_k^{(m+1)}=\frac{1}{n_k^{(m)}}\sum_{i=0}^{N-1}\gamma_{ik}^{(m)}\boldsymbol{y}_i$$

$$\mathbf{C}_{\mathbf{k}}^{(m+1)}=\frac{1}{n_k^{(m)}}\sum_{i=0}^{N-1}\gamma_{ik}^{(m)}\left(\boldsymbol{y}_i-\boldsymbol{\mu}_k\right)\left(\boldsymbol{y}_i-\boldsymbol{\mu}_k\right)^T$$

S4 收敛性验证，计算 $l(\theta^{(m+1)} | y)$ 并检查下式

$$\left\| l(\theta^{(m+1)} | y) - l(\theta^{(m)} | y) \right\| < \varepsilon$$

若成立则停止，否则 $m = m + 1$ 转 S2。

GMM参数估计的EM算法实例

Old Faithful 数据集