高斯混合模型

$$P(X)=\sum_{k=1}^{K}\pi_k\,N(X\mid u_k,\Sigma_k)$$

$$\begin{cases} N(X\mid u_k,\Sigma_k)=\dfrac{1}{\sqrt{(2\pi)^d|\Sigma_k|}}\exp\left\{-\frac{1}{2}(X-u_k)^T\Sigma_k^{-1}(X-u_k)\right\} \\ \sum_{k=1}^{K}\pi_k=1 \end{cases}$$

用极大似然法估计概率密度

输入：$\{X_i\mid i=1\sim N\}$

最小化：$E(\{\pi_k,u_k,\Sigma_k\mid k=1\sim K\})$

$$=-\sum_{i=1}^{N}\log[P(X_i)]$$

$$=-\sum_{i=1}^{N}\log\left[\sum_{k=1}^{K}\pi_k\cdot\frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}}\exp\left\{-\frac{1}{2}(X_i-u_k)^T\Sigma_k^{-1}(X_i-u_k)\right\}\right]$$

非凸问题，无法求全局极值，只能求局部极值

↓

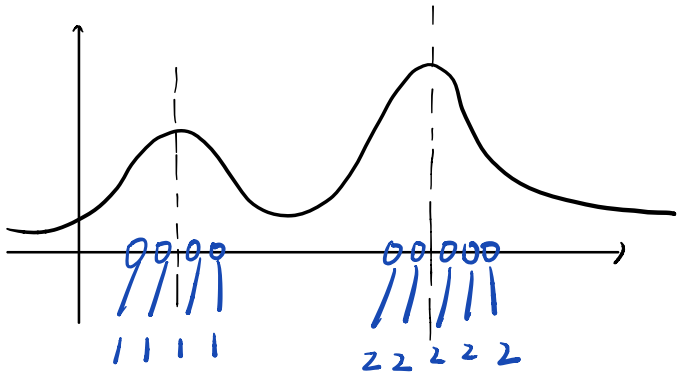
① 梯度下降

② 启发式方法

③ EM算法

1. 不需调参一定收敛
2. 编程简单
3. 理论优美

EM算法：



随机化所有 $\{\pi_k, \mu_k, \Sigma_k\}$ $k=1\sim K$

$k=\{1,2\}$

高斯混合模型EM（Expectation-Maximization）

① 随机化 $\{\pi_k, \mu_k, \Sigma_k\}$ $k=1\sim K$

② E-Step：

$$\gamma_{nk} = \frac{\pi_k \cdot N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)} \quad (n=1\sim N) \quad (k=1\sim K)$$

（第n个样本落在第k个高斯的概率）

③ 更新：M-Step

$$N_k = \sum_{n=1}^{N}\gamma_{nk}$$ （所有N个样本中有多少属于第k个样本）

$$\pi_k^{(new)} = \frac{N_k}{N}$$ （π_k：第k个高斯的概率）

$$\mu_k^{(new)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\cdot x_n$$ （μ_k：第k个高斯的均值）

$$\Sigma_k^{(new)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}(x_n-\mu_k^{(new)})(x_n-\mu_k^{(new)})^T$$

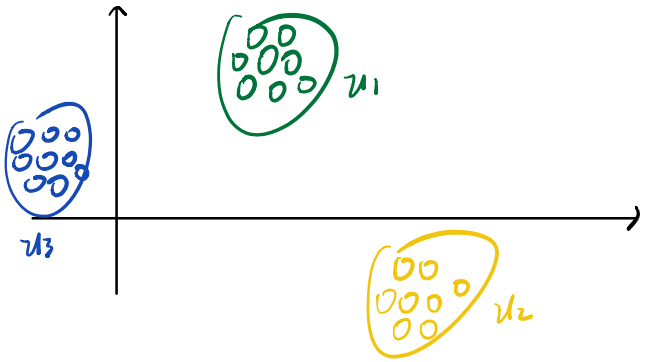
$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(x_n-u_k)(x_n-u_k)^T$$

(Σ_k: 第k个高斯的协方差矩阵)

④ 回到② 直到收敛

(应用: 高斯混合模型在说话人识别方面的应用)

K-均值聚类 (K-means clustering)



(如何让机器自动聚类?)

$$E(\{u_k\}) = \sum_{k=1}^{K}\sum_{n=1}^{N} \|x_i-u_k\|^2$$

(第②③都使E变小, 而E有下界0 一定收敛)

问题: 输入N个样本 $\{x_i\}$ $i=1\sim N$

输出N个样本类别 $\{z_i\}$ $i=1\sim N$

其中 $z_i = 1, 2, 3, \ldots, K$

① 随机化 $u_1, \ldots, u_k$

② E-Step:

$$z_i = \arg\min_k \|x_i - u_k\| \quad i=1\sim N$$

(离谁近属于谁)

③ M-Step：

$$N_k = \sum_{i=1}^{N} \mathbb{I}(z_i = k)$$ （N个样本中有多少属于第k类）

$$u_k = \frac{1}{N_k} \sum_{\substack{i=1 \\ z_i=k}}^{N} x_i$$ （第k类样本的均值）

④ 回到② 直到收敛

（应用：基于K-均值聚类的图像矢量化/压缩图形）