中国科学院南京分院
Nanjing Branch of Chinese Academy of Sciences

人工智能原理与算法

7. 聚类算法

夏睿

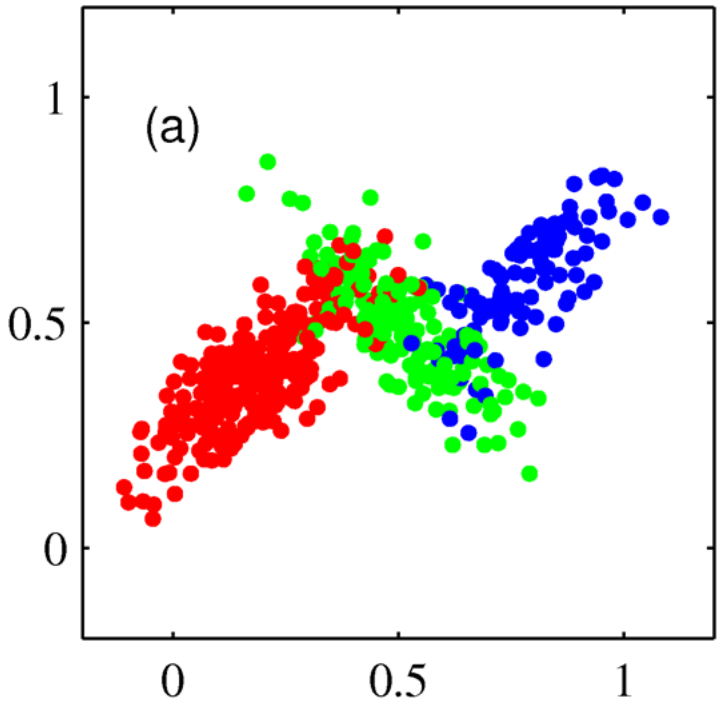
2023.3.22

目录

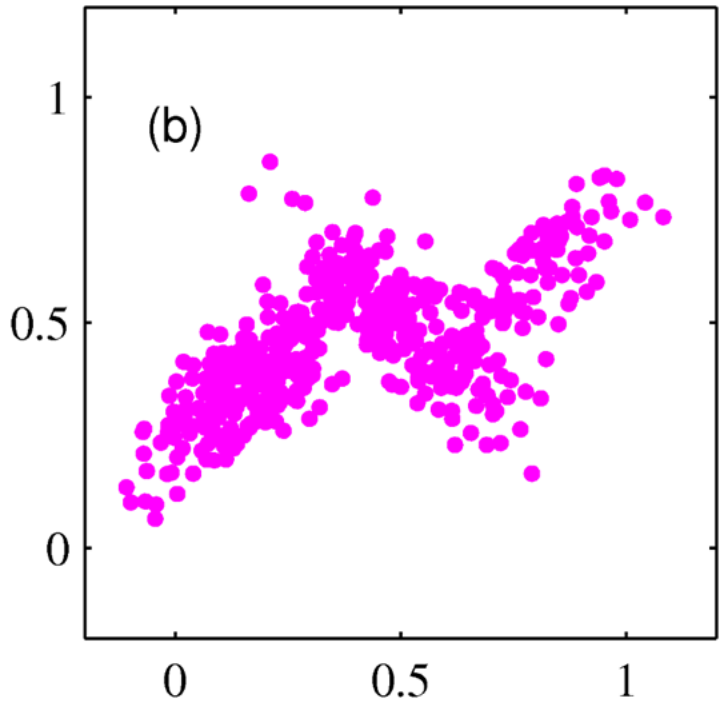
- 聚类 vs. 分类
- K-均值聚类算法
- 高斯混合模型
- 层次聚类算法
- 聚类任务的性能评估

分类 vs. 聚类

监督学习
分类



无监督学习
聚类

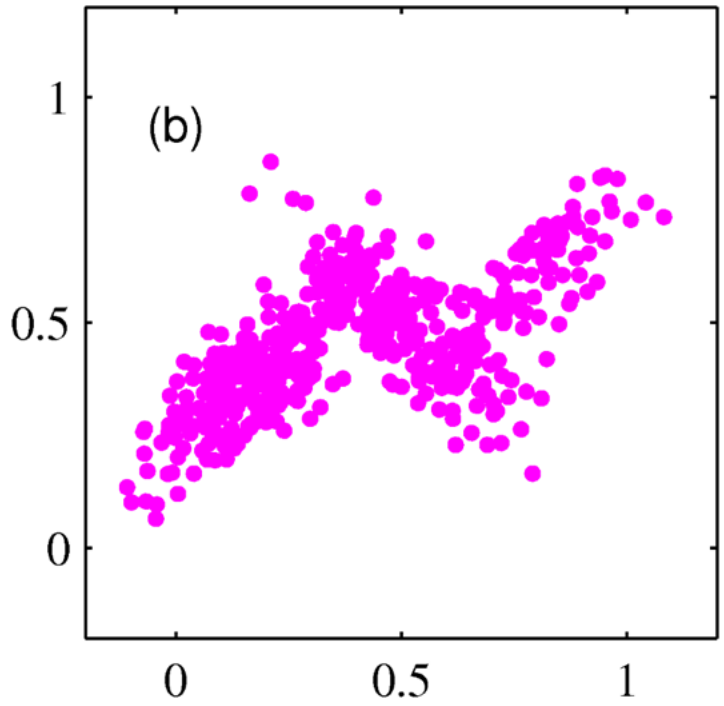


聚类算法的类型

- 基于分区的算法（K-均值聚类等）
- 基于分布的算法（高斯混合模型）
- 基于层次的算法（层次聚类）
- 基于密度的算法（DBSCAN等）
- 基于图论的算法（谱聚类等）
- 基于网格的算法（CLIQUE聚类等）
- ……

无监督学习

聚类



K-均值（K-Means）聚类

K-均值算法的历史

- K-均值（K-means）是一种基于分区的聚类算法，也是使用最为广泛的聚类算法。
- K-均值标准算法作为一种脉冲编码调制技术由Bell Labs的Stuart Lloyd于1957年首次提出，然而直到1982年它才作为期刊论文发表。1965年，Edward W. Forgy发表了本质相同的方法。后人也将K-均值标准算法称为Lloyd-Forgy算法。
- James MacQueen于1967年首次使用了“k-means”一词。
- 它有很多变体，如Hartigan-Wong算法、k-medoids聚类、k-medians聚类、高斯混合模型，等等。

学习目标

- 对于给定数据集$\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(N)}\}$，$K$-均值聚类的目标是将$N$个样本分成$K$个聚类，以最小化各个聚类内的平方距离之和，称为聚类内平方和（within-cluster sum of squares, WCSS）：

$$\arg\min_{\boldsymbol{C}} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} \|\boldsymbol{x} - \boldsymbol{m}_k\|^2$$

- 为最小化WCSS，标准的K-均值算法（即Lloyd-Forgy方法）使用了迭代优化方法：
  - 在每个迭代步中，先计算每个样本与K个聚类中心（即均值）之间的距离。
  - 然后将样本分配给最近的中心点所属的聚类，并更新现有聚类的中心。
  - 重复此过程，直到算法收敛。

迭代优化

- 初始化：从数据集中选择K个样本作为初始均值（也可在样本点区域附近随机分配）。
- **分配：**将每个样本分配给最近的聚类，即平方欧氏距离最小：

$$C_i^{(t)} = \left\{ \boldsymbol{x}: d\left(\boldsymbol{x}, \boldsymbol{m}_i^{(t)}\right) \le d\left(\boldsymbol{x}, \boldsymbol{m}_j^{(t)}\right), j = 1, \cdots, K \right\}$$

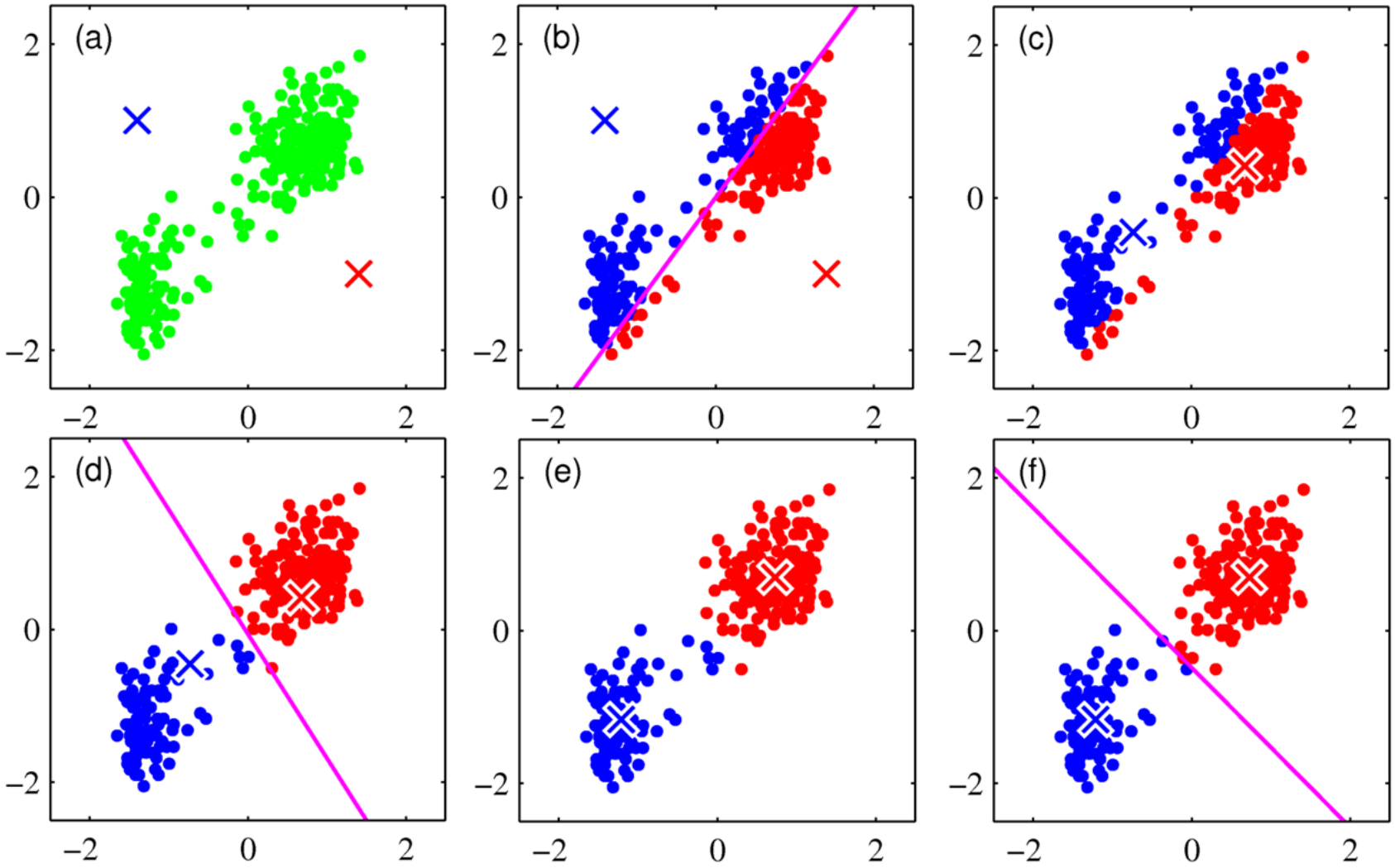
其中$d\left(\boldsymbol{x}, \boldsymbol{m}_k^{(t)}\right) = \left\| \boldsymbol{x} - \boldsymbol{m}_k^{(t)} \right\|^2$，$t$表示迭代步。

- **更新：** 更新每个聚类的均值：

$$\boldsymbol{m}_i^{(t+1)} = \frac{1}{\left| C_i^{(t)} \right|} \sum_{\boldsymbol{x}_j \in C_i^{(t)}} \boldsymbol{x}_j$$

- 重复以上两个步骤，直到算法收敛到局部最小值（当分配不再改变）。

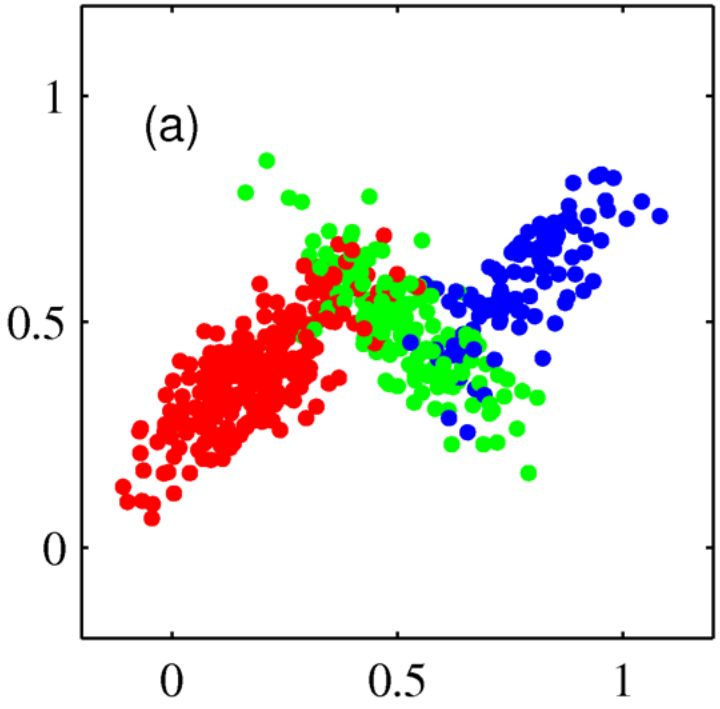
示例



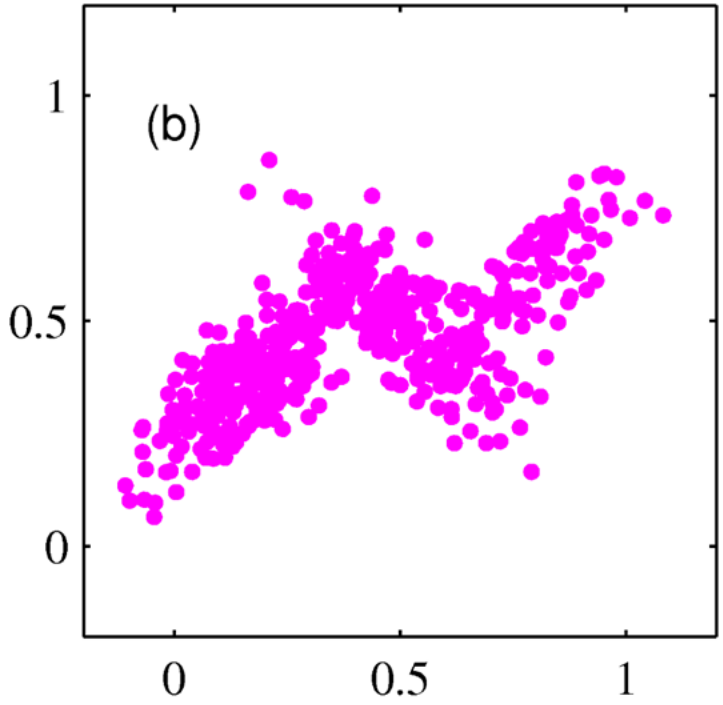
高斯混合模型
(Gaussian Mixture Model, GMM)

生成式模型（监督 vs. 无监督）

分类（监督学习）



聚类(无监督学习)



$$p(\boldsymbol{x}, y = j) = p(y = j)p(\boldsymbol{x}|y = j)$$

$$p(\boldsymbol{x}) = \sum_{j} p(\boldsymbol{x}, y = j)$$

模型假设

- 联合分布（监督学习）

$$p(\boldsymbol{x}, y=j)=p(y=j)p(\boldsymbol{x}|y=j)=\pi_j\frac{1}{\sqrt{(2\pi)^M|\boldsymbol{\Sigma}_j|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-u_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-u_j)\right)$$

- 边缘分布（无监督学习）

$$p(\boldsymbol{x})=\sum_j p(\boldsymbol{x}, y=j)=\sum_j \pi_j\frac{1}{\sqrt{(2\pi)^M|\boldsymbol{\Sigma}_j|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-u_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-u_j)\right)$$

最大似然估计（监督学习）

- 联合分布的似然函数

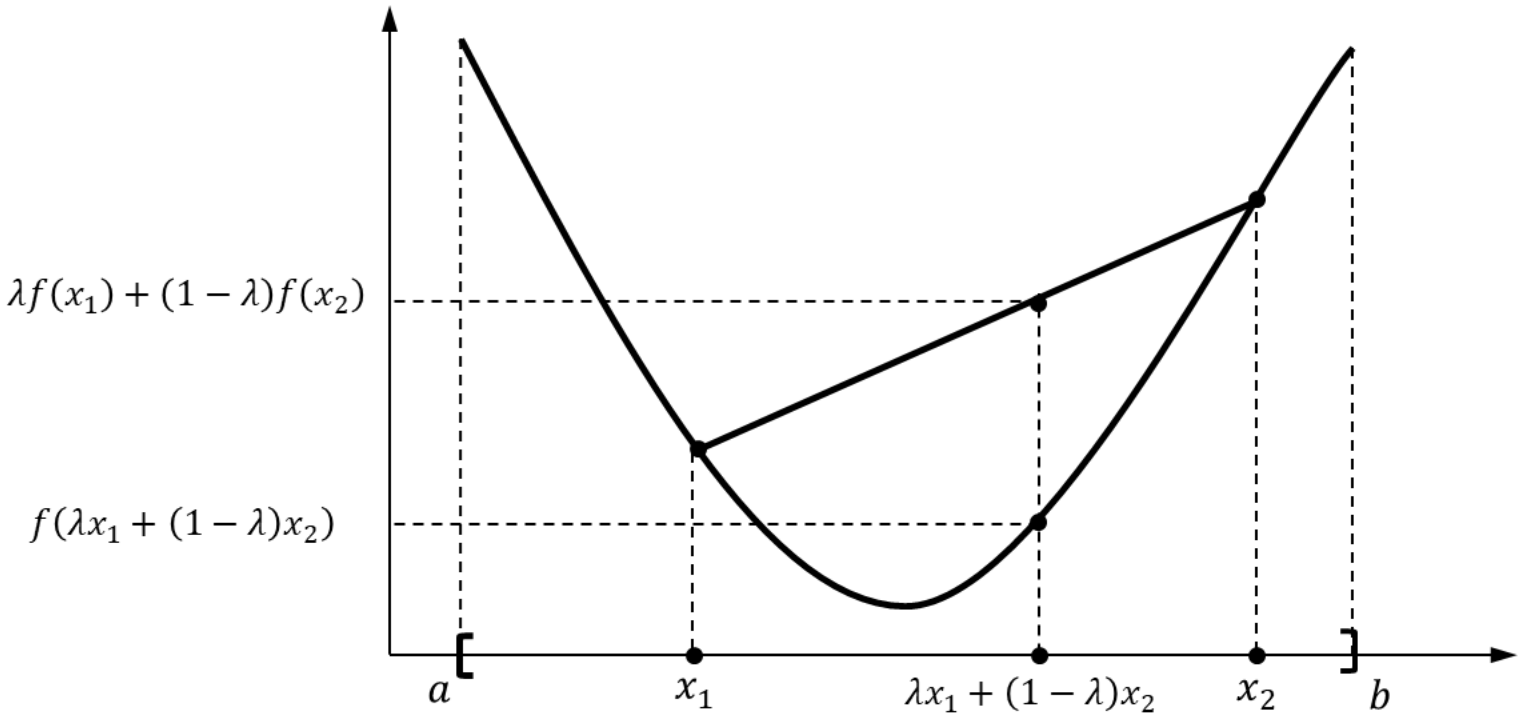
$$L = \sum_{k=1}^{N} \log p\left(\boldsymbol{x}^{(k)}, y^{(k)} \middle| \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$
$$= \sum_{k=1}^{N} \sum_{j=1}^{C} I\left(y^{(k)} = j\right) \log \pi_j \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}^{(k)} - \boldsymbol{u}_j\right)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} \left(\boldsymbol{x}^{(k)} - \boldsymbol{u}_j\right)\right)$$

- 最大似然估计解析解

$$\pi_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)}{N}$$

$$\boldsymbol{\mu}_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)\boldsymbol{x}^{(k)}}{\sum_{k=1}^{N} I(y^{(k)} = j)} \qquad \boldsymbol{\Sigma}_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)^{\mathrm{T}}}{\sum_{k=1}^{N} I(y^{(k)} = j)}$$

Jensen不等式



f是凸函数 ⇒ $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$

⇓

$$f\left(\sum_{j=1}^{n} \lambda_i x_i\right) \le \sum_{i=1}^{n} \lambda_i f(x_i) \quad \left(\lambda_i \ge 0\,, \sum_{i=1}^{n} \lambda_i = 1\right)$$

最大似然估计（无监督学习）

- 边缘分布的似然函数

log sum问题，难以直接进行最大似然估计

$$L(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log p\left(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}\right) = \sum_{k=1}^{N} \log \sum_{j=1}^{C} p\left(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta}\right)$$

$$= \sum_{k=1}^{N} \log \sum_{j=1}^{C} q(y^{(k)} = j) \frac{p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta})}{q(y^{(k)} = j)}$$

$$\geq \sum_{k=1}^{N} \sum_{j=1}^{C} q(y^{(k)} = j) \log \frac{p\left(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta}\right)}{q(y^{(k)} = j)}$$

考虑$y^{(k)}$为随机变量，$q(y^{(k)} = j)$为其分布；$g(y^{(k)}) = \frac{p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \theta)}{q(y^{(k)} = j)}$是关于$y^{(k)}$的函数

log为凹函数，所以：$\log\left(E_{y^{(k)} \sim q} g(y^{(k)})\right) \geq E_{y^{(k)} \sim q}\left(\log g(y^{(k)})\right)$

期望步（Expectation Step）

- 似然函数的下界

sum log形式，便于进行最大似然估计

$$L(\boldsymbol{\theta}) \geq \tilde{L}(\boldsymbol{\theta}, q) = \sum_{k=1}^{N} \sum_{j=1}^{C} q(y^{(k)} = j) \log \frac{p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta})}{q(y^{(k)} = j)}$$

- 等号成立条件

$$\frac{p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta}^{old})}{q(y^{(k)} = j)} = const \quad \Longleftrightarrow \quad q(y^{(k)} = j) = \frac{p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta}^{old})}{\sum_{j=1}^{C} p(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\theta}^{old})}$$

$$= p(y^{(k)} = j | \boldsymbol{x}^{(k)}; \boldsymbol{\theta}^{old})$$

E-step：对于确定的$\boldsymbol{\theta}^{old}$，取$q^{old}(y^{(k)} = j) = p(y^{(k)} = j | \boldsymbol{x}^{(k)}; \boldsymbol{\theta}^{old})$，使得在$\boldsymbol{\theta}^{old}$处，$\tilde{L}(\boldsymbol{\theta}^{old}, q) = L(\boldsymbol{\theta}^{old})$

最大化步（Maximization Step）

- 无监督学习似然函数下界的最大化

$$\arg\max_{\boldsymbol{\theta}} \tilde{L}(\boldsymbol{\theta}, Q) = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{N}\sum_{j=1}^{C} p\left(y^{(k)} = j|\boldsymbol{x}^{(k)}; \boldsymbol{\theta}^{old}\right) \log \frac{p\left(x^{(k)}, y^{(k)}; \boldsymbol{\theta}\right)}{p\left(y^{(k)}|\boldsymbol{x}^{(k)}; \boldsymbol{\theta}^{old}\right)}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{N}\sum_{j=1}^{C} p\left(y^{(k)} = j|\boldsymbol{x}^{(k)}; \boldsymbol{\theta}^{old}\right) \log p\left(x^{(k)}, y^{(k)}; \boldsymbol{\theta}\right)$$

M-step：进一步调节$\boldsymbol{\theta}$，最大化似然函数的下界$\tilde{L}\left(\boldsymbol{\theta}, q^{old}\right)$，得到下一步参数$\boldsymbol{\theta}^{new}$

- 回顾监督学习的似然函数

两者具有极其相似的形式，监督下为标注的独热分布，无监督下为预测的后验分布

$$L_s(\boldsymbol{\theta}) = \log \prod_{k=1}^{N} p\left(\boldsymbol{x}^{(k)}, y^{(k)}; \boldsymbol{\theta}\right) = \sum_{k=1}^{N}\sum_{j=1}^{C} I\left(y^{(k)} = j\right) \log p\left(x^{(k)}, y^{(k)}; \boldsymbol{\theta}\right)$$

期望最大化算法（EM Algorithm）训练过程

- 初始化参数： $\pi_j^{(0)}, \boldsymbol{\mu}_j^{(0)}, \boldsymbol{\Sigma}_j^{(0)} (j = 1, \cdots, C)$
- E-step　　基于第t步参数，预测未标注数据的后验概率，作为第t步q分布

$$q^{(t)}\left(y^{(k)} = j\right) = p\left(y^{(k)} = j \middle| \boldsymbol{x}^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right) = \frac{p\left(y^{(k)} = j; \pi_j^{(t)}\right) p(\boldsymbol{x}^{(k)} | y^{(k)} = j; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\sum_{j'=1}^{C} p\left(y^{(k)} = j'; \pi_{j'}^{(t)}\right) p(\boldsymbol{x}^{(k)} | y^{(k)} = j'; \boldsymbol{\mu}_{j'}^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}$$

- M-step　　基于第t步q分布，最大化似然函数下界，获得第t+1步参数

$$\arg\max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \tilde{L}\left(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, Q^{(t)}\right) = \sum_{k=1}^{N} \sum_{j=1}^{C} p\left(y^{(k)} = j \middle| \boldsymbol{x}^{(k)}; \boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\right) \log p\left(\boldsymbol{x}^{(k)}, y^{(k)} = j; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) + const$$

- 重复E-step、M-step，直至模型收敛

参数估计结果（监督 vs. 无监督）

监督学习

$$\pi_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)}{N}$$

$$\boldsymbol{\mu}_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)\boldsymbol{x}^{(k)}}{\sum_{k=1}^{N} I(y^{(k)} = j)}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{k=1}^{N} I(y^{(k)} = j)(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_j)^T}{\sum_{k=1}^{N} I(y^{(k)} = j)}$$

无监督学习(M-step第t+1迭代)

$$\pi_j^{(t+1)} = \frac{\sum_{k=1}^{N} p(y^{(k)} = j|x^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{N}$$

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{k=1}^{N} p(y^{(k)} = j|\boldsymbol{x}^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})\boldsymbol{x}^{(k)}}{\sum_{k=1}^{N} p(y^{(k)} = j|x^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{k=1}^{N} p(y^{(k)} = j|\boldsymbol{x}^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_j)^T)}{\sum_{k=1}^{N} p(y^{(k)} = j|x^{(k)}; \pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}$$

参数估计：每个样本从监督学习的绝对类别，转化为无监督下的类别概率。

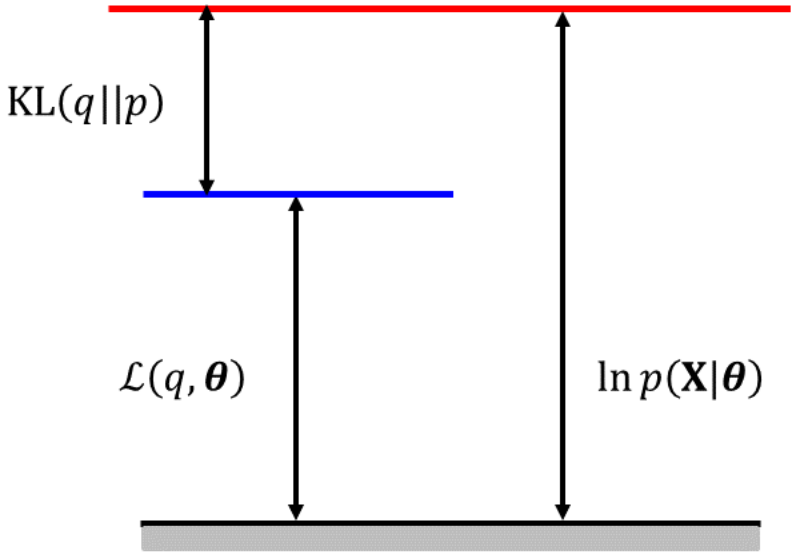
从K-L散度视角的理解

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{p(\boldsymbol{x};\boldsymbol{\theta}) q\left(y=j;\boldsymbol{\theta}^{old}\right) p(\boldsymbol{x}, y=j;\boldsymbol{\theta})}{q(y=j;\boldsymbol{\theta}^{old}) p(\boldsymbol{x}, y=j;\boldsymbol{\theta})}$$

$$= \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{p(\boldsymbol{x}, y;\boldsymbol{\theta})}{q(y=j;\boldsymbol{\theta}^{old})} + \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{q\left(y=j;\boldsymbol{\theta}^{old}\right)}{p(y=j|\boldsymbol{x};\boldsymbol{\theta})}$$

$$= \tilde{L}\left(\boldsymbol{\theta}, q\left(y=j;\boldsymbol{\theta}^{old}\right)\right) + \mathrm{KL}\left(q(y=j;\boldsymbol{\theta}^{old}) || p(y|\boldsymbol{x};\boldsymbol{\theta})\right)$$

似然 = 下界 + KL距离



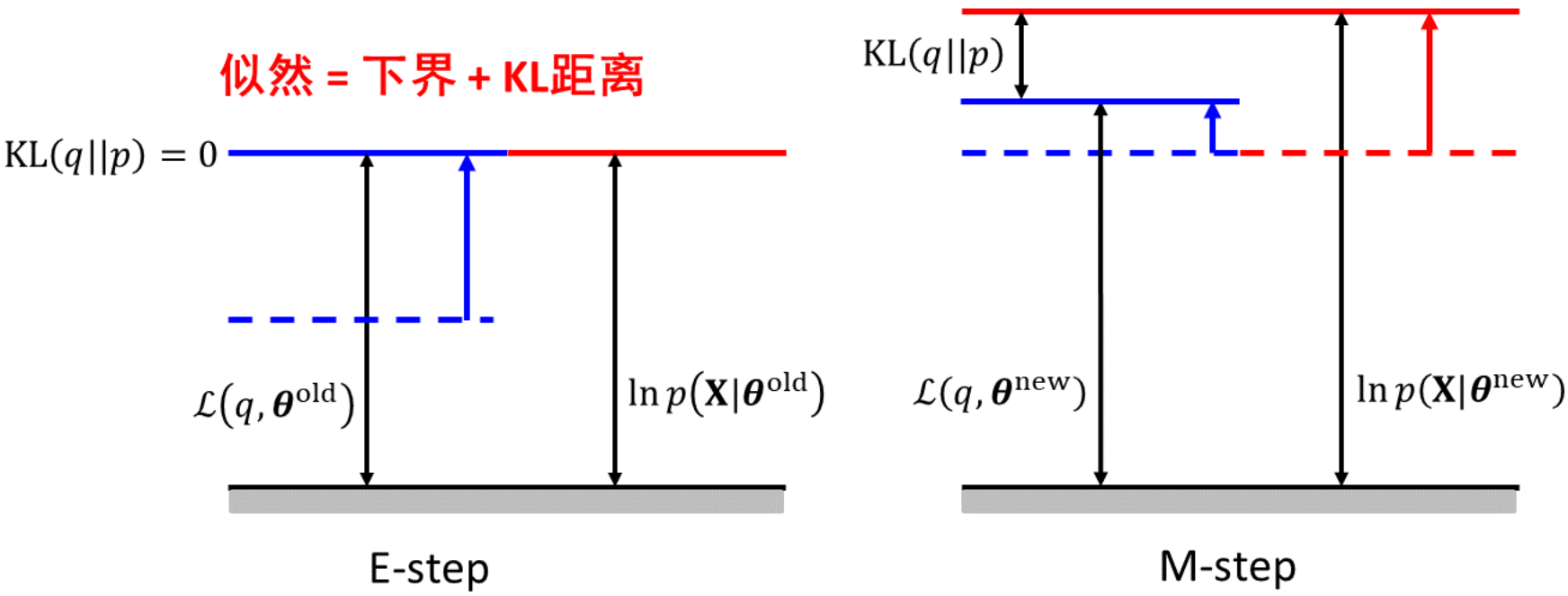
从K-L散度视角的理解

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{p(\boldsymbol{x};\boldsymbol{\theta}) q\left(y=j;\boldsymbol{\theta}^{old}\right) p(\boldsymbol{x}, y=j;\boldsymbol{\theta})}{q(y=j;\boldsymbol{\theta}^{old}) p(\boldsymbol{x}, y=j;\boldsymbol{\theta})}$$

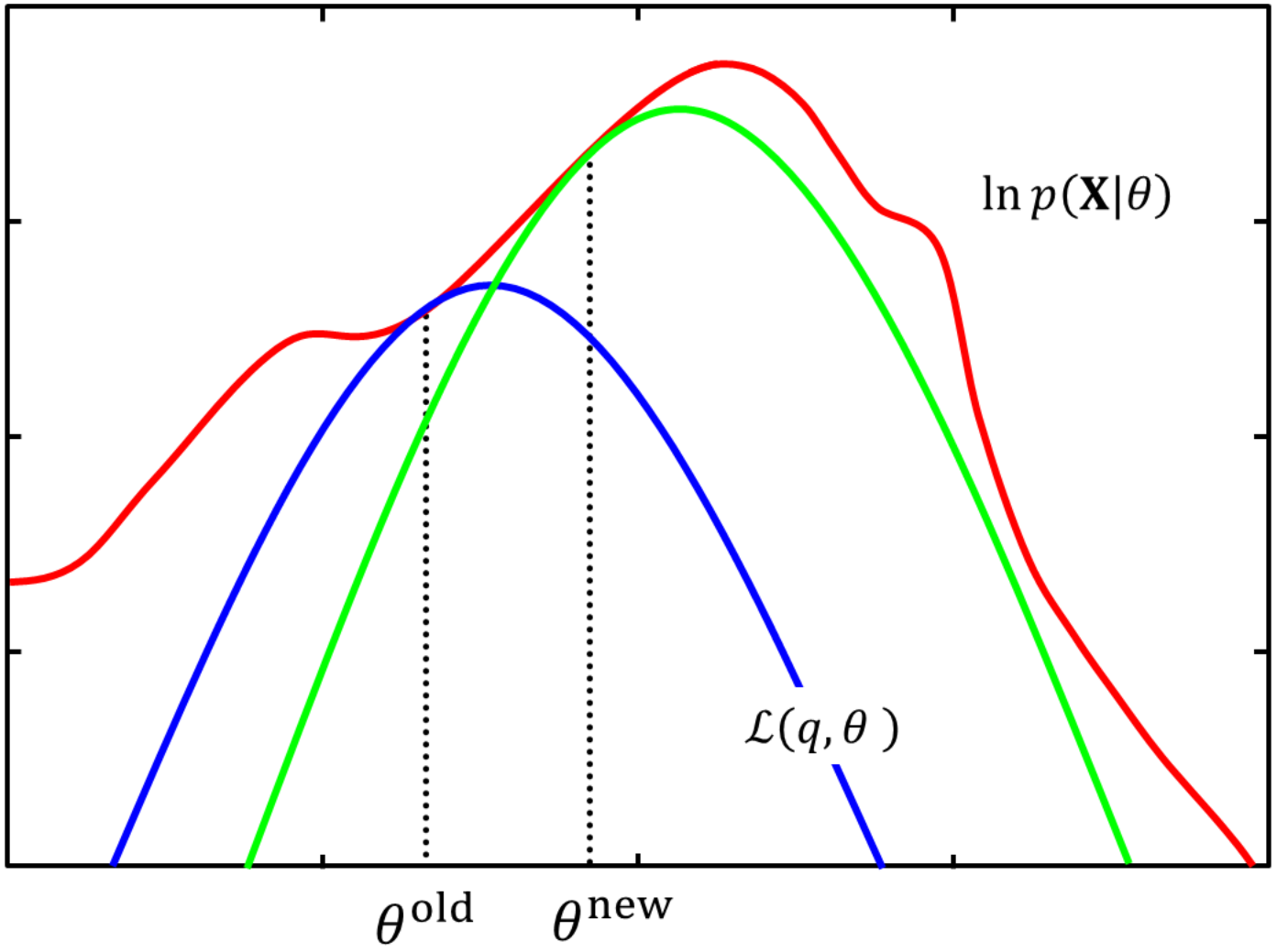
$$= \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{p(\boldsymbol{x}, y;\boldsymbol{\theta})}{q(y=j;\boldsymbol{\theta}^{old})} + \sum_{j=1}^{C} q(y=j;\boldsymbol{\theta}^{old}) \log \frac{q\left(y=j;\boldsymbol{\theta}^{old}\right)}{p(y=j|\boldsymbol{x};\boldsymbol{\theta})}$$

$$= \tilde{L}\left(\boldsymbol{\theta}, q\left(y=j;\boldsymbol{\theta}^{old}\right)\right) + \mathrm{KL}\left(q(y=j;\boldsymbol{\theta}^{old}) || p(y|\boldsymbol{x};\boldsymbol{\theta})\right)$$

似然 = 下界 + KL距离



EM算法迭代寻优图释


$\ln p(\mathbf{X}|\theta)$
$\mathcal{L}(q,\theta)$
θ^{old}
θ^{new}

K-Means与GMM的对比

K-Means vs. GMM

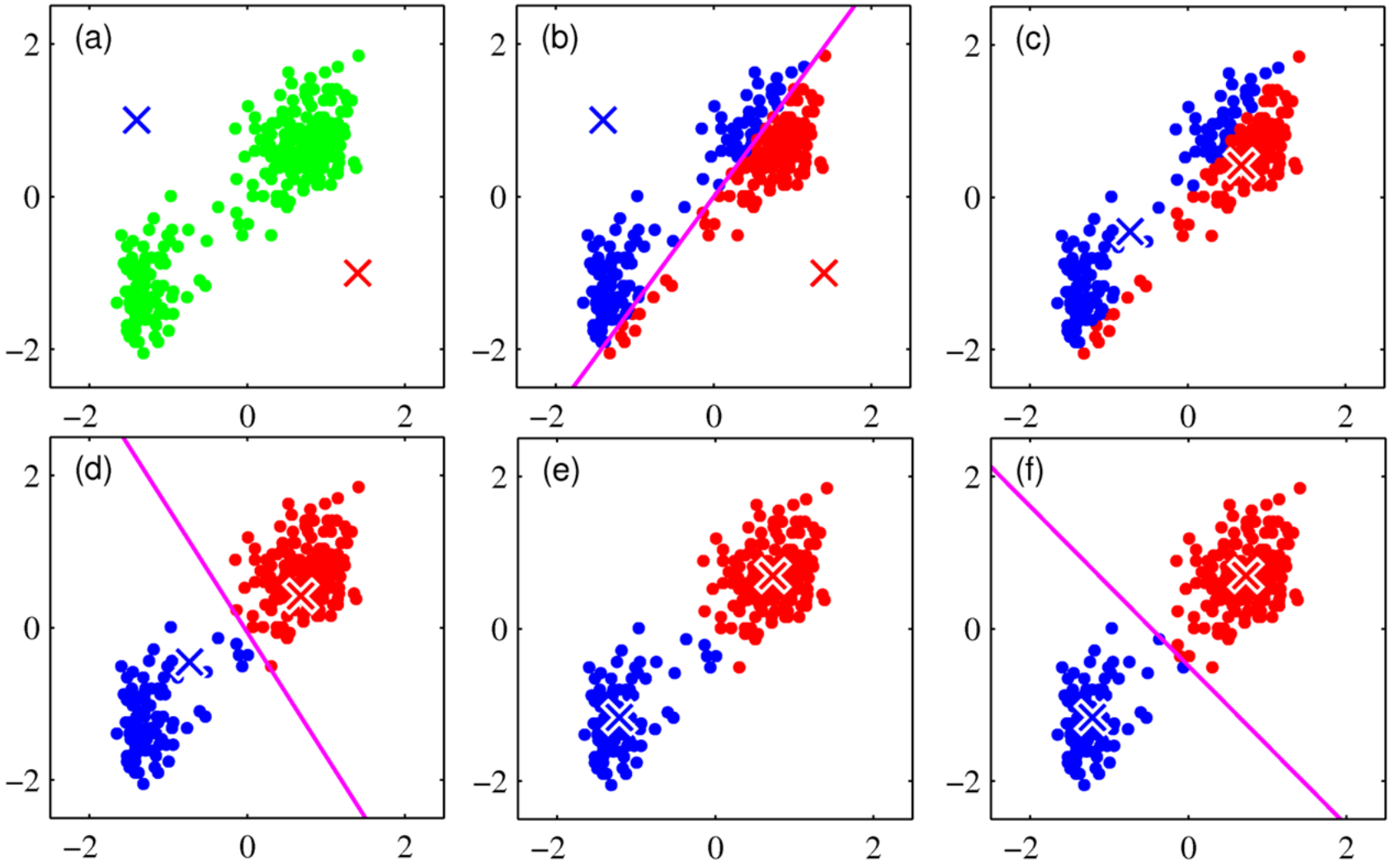
K-Means

- 初始化：K个初始聚类中心。
- **分配：**基于第t步聚类中心，计算样本到每个聚类中心距离，将样本分配给最近的聚类；
- **更新：** 根据样本分类的结果，更新得到第t+1步的聚类均值；
- 重复以上两个步骤，直到算法收敛。

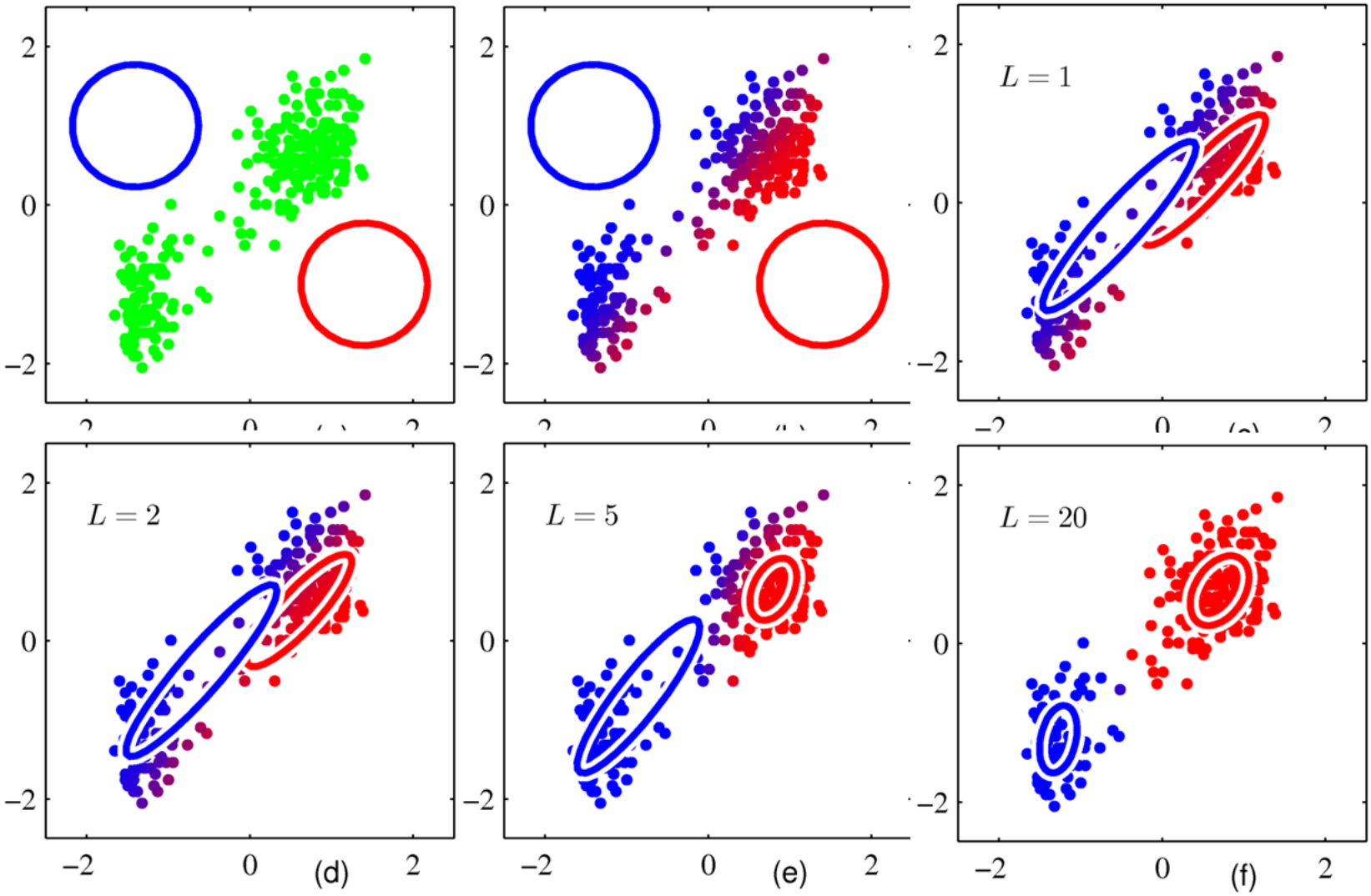
GMM

- 初始化：K个聚类，初始化参数 $\pi_j^{(0)}, \boldsymbol{\mu}_j^{(0)}, \boldsymbol{\Sigma}_j^{(0)} (j = 1, \cdots, K)$
- **E-step：**基于第t步参数，计算每个样本属于各类的后验概率，作为第t步q分布；
- **M-step：**基于第t步q分布，最大化似然函数下界，得到第t+1步参数；
- 重复以上两个步骤，直到算法收敛。

K均值算法示例



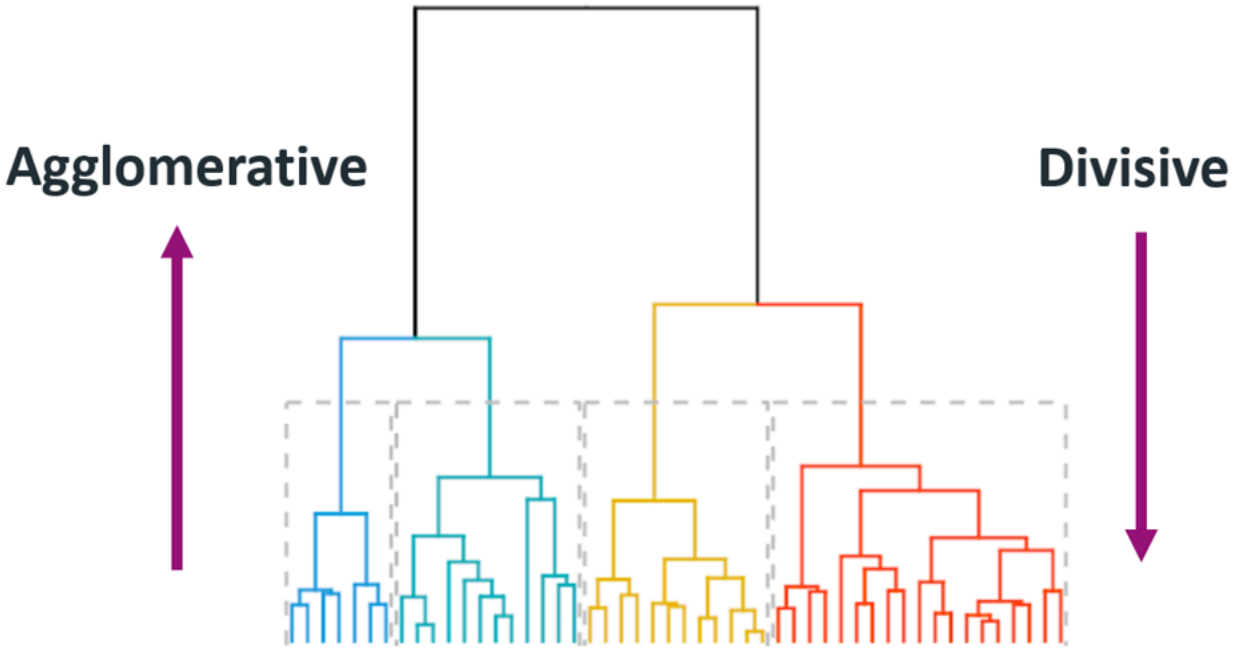
GMM算法示例



层次聚类算法

层次聚类

- 依据一种层次架构将数据逐层进行聚合或分裂，最终将数据对象组织成一棵聚类树状的结构
- 按照聚类树生成的方式
  - 自底向上的聚合式层次聚类（agglomerative hierarchical clustering）
  - 自顶向下的分裂式层次聚类（divisive hierarchical clustering）



样本的相似性度量

- 欧氏距离（Euclidean distance）

$$d(\boldsymbol{a},\boldsymbol{b})=\left(\sum_{i=1}^{M}(a_i-b_i)^2\right)^{1/2}$$

- 曼哈顿距离（Manhattan Distance）

$$d(\boldsymbol{a},\boldsymbol{b})=\sum_{i=1}^{M}|a_i-b_i|$$

- 切比雪夫距离（Chebyshev Distance）

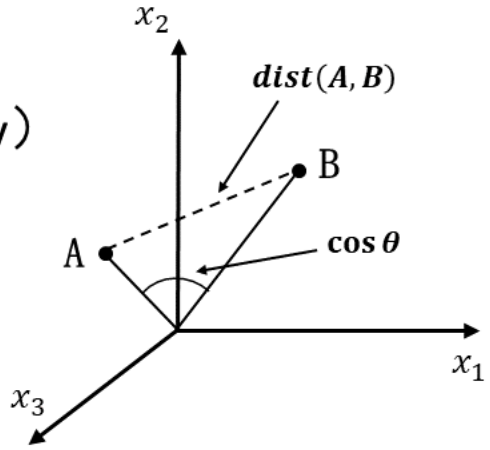
$$d(\boldsymbol{a},\boldsymbol{b})=\max_i|a_i-b_i|$$

- 闵可夫斯基距离（Minkowski Distance）

$$d(\boldsymbol{a},\boldsymbol{b})=\left(\sum_{i=1}^{M}|a_i-b_i|^p\right)^{1/p}$$

- 余弦相似性（Cosine Similarity）

$$Cos(\boldsymbol{a},\boldsymbol{b})=\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$$



取值范围：$[-1,1]$，
夹角越小取值越大

聚类簇的相似性度量

- 最短距离法（Single Linkage）

$$d(C_m, C_n) = \min_{\boldsymbol{x}^{(i)} \in C_m, \boldsymbol{x}^{(j)} \in C_n} d\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)$$

- 最长距离法（Complete Linkage）

$$d(C_m, C_n) = \max_{\boldsymbol{x}^{(i)} \in C_m, \boldsymbol{x}^{(j)} \in C_n} d\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)$$

- 簇平均法（Average Linkage）

$$d(C_m, C_n) = \frac{1}{|C_m| \cdot |C_n|} \sum_{\boldsymbol{x}^{(i)} \in C_m} \sum_{\boldsymbol{x}^{(j)} \in C_n} d\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)$$

- 重心法

$$d(C_m, C_n) = d(\overline{\boldsymbol{x}}(C_m), \overline{\boldsymbol{x}}(C_n))$$

- 离差平方和法（Ward's method）

$$d(C_m, C_n) = \sum_{\boldsymbol{x}^{(k)} \in C_m \cup C_n} d\left(\boldsymbol{x}^{(k)}, \overline{\boldsymbol{x}}(C_m \cup C_n)\right) - \sum_{\boldsymbol{x}^{(i)} \in C_m} d\left(\boldsymbol{x}^{(i)}, \overline{\boldsymbol{x}}(C_m)\right) - \sum_{\boldsymbol{x}^{(j)} \in C_n} d\left(\boldsymbol{x}^{(j)}, \overline{\boldsymbol{x}}(C_n)\right)$$

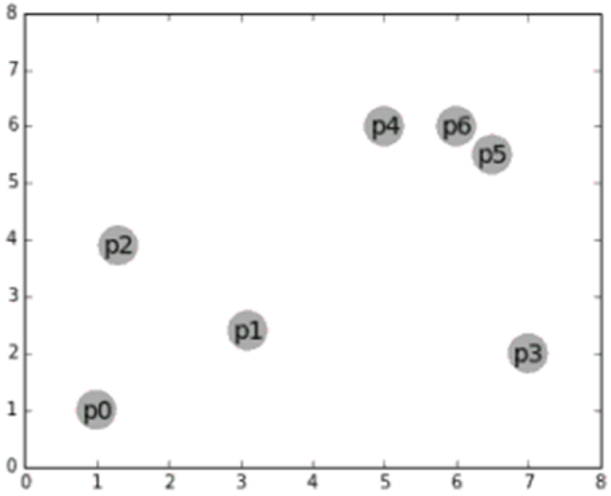
自底向上的聚合式层次聚类流程

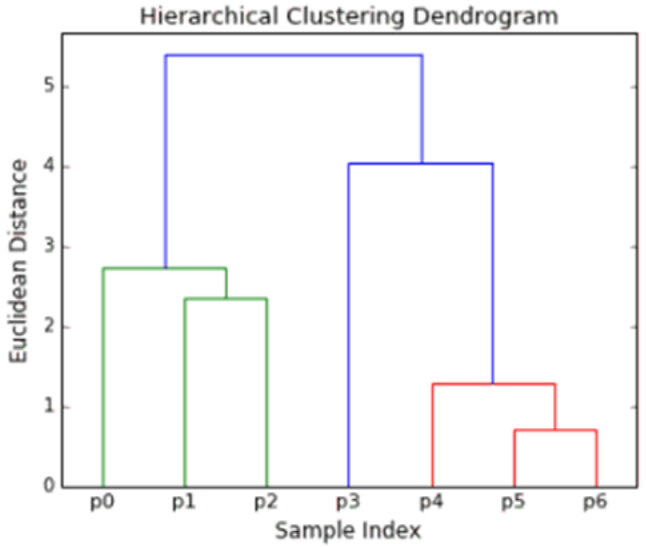
输入：数据集$\mathcal{D}=\{\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\dots,\boldsymbol{x}^{(N)}\}$，聚类簇数为$K$；

输出：聚类划分$\mathcal{C}=\{C_1,C_2,\dots,C_K\}$。

算法：

1. for $i=1,\dots,N$
2. $\quad C_i=\{\boldsymbol{x}^{(i)}\}$
3. for $i=1,\dots,N$
4. $\quad$ for $j=1,\dots,N$
5. $\quad\quad$ 计算两两簇间的相似性$d(C_i,C_j)$
6. while $size(\mathcal{C})>K$
7. $\quad$ 查找距离最近的两个簇C_{i^*}和C_{j^*}
8. $\quad$ for $h=1,\dots,size(\{C_k\})$
9. $\quad\quad$ if $h\neq i^*$ and $h\neq j^*$
10. $\quad\quad\quad$ 更新簇间相似度$d\left(C_h,C_{i^*}\cup C_{j^*}\right)$
11. $\quad$ 簇集合$\mathcal{C}$中删除C_{i^*}和C_{j^*}
12. $\quad$ 簇集合$\mathcal{C}$中添加$C_{i^*}\cup C_{j^*}$
13. $\quad$ 更新$\mathcal{C}$中各簇标号，记录各簇样本标号



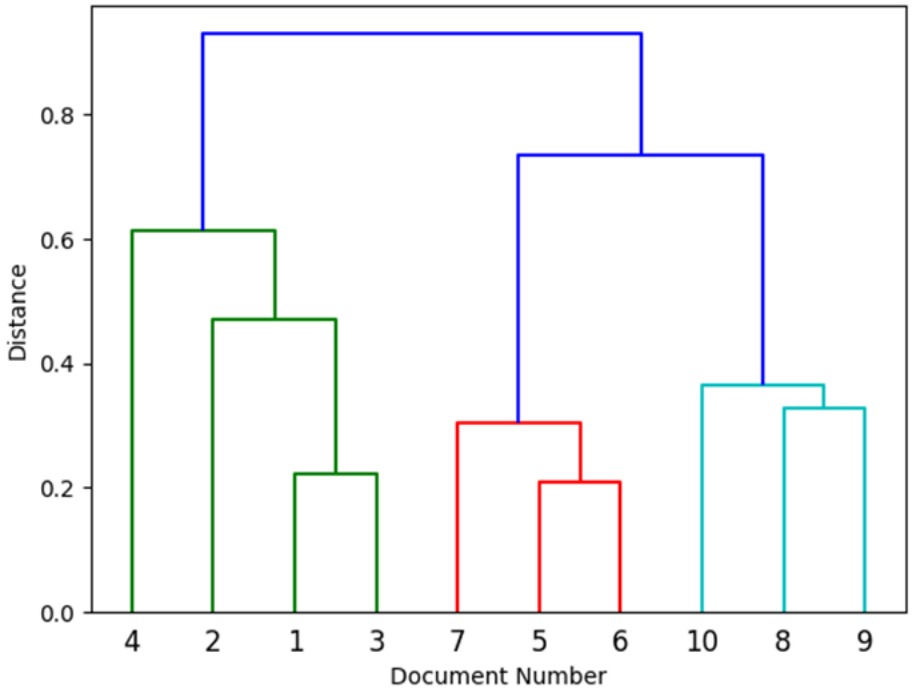


示例

ID	文本
1	北京 理工 大学 计算机 专业 创建 于 1958年 是 中国 最早 设立 计算机 专业 的 高校 之一
2	北京 理工 大学 学子 在 第四 届 中国 计算机 博弈 锦标赛 中 夺冠
3	北京 理工 大学 体育馆 是 2008 年 中国 北京 奥林匹克 运动会 的 排球 预赛 场地
4	第五 届 东亚 运动会 中国 军团 奖牌 总数 创 新高 男女 排球 双双 夺冠
5	人工 智能 也 称 机器 智能 是 指 由 人工 制造 出來 的 系统 所 表现 出来 的 智能
6	人工 智能 是 计算机 科学 的 一个 分支 它 企图 生产 出 一种 能 以 人类 智能 相似 的 方式 做出 反应 的 智能 机器
7	AlphaGo 人工 智能 对决 围棋 世界 冠军 柯洁 的 三场 赛事 以 人类 完败 结果 告终
8	曲曲折折 的 荷塘 上面 弥望的 是 田田 的 叶子 叶子 出水 很高 像 亭亭 的 舞女 的裙
9	月光 如 流水 一般 静静 地 泻 在 这 一片 叶子 和 花上 薄薄 的 青雾 浮起 在 荷塘 里
10	叶子 底下 是 脉脉 的 流水 遮住 了 不能 见 一些 颜色 而 叶子却 更见 风致 了

示例

ID	降维后的文本
1	北京 理工 大学 计算机 专业 年 是 中国 的
2	北京 理工 大学 在 届 中国 计算机 夺冠
3	北京 理工 大学 是 年 中国 北京 运动会 的 排球
4	届 运动会 中国 排球 夺冠
5	人工 智能 机器 是 的
6	人工 智能 是 计算机 的 以 人类 机器
7	人工 智能 的 以 人类
8	的 荷塘 是 叶子
9	流水 在 叶子 的 荷塘
10	叶子 是 的 流水 了



聚类性能评估
（聚类有效性分析）

聚类性能评估（外部标准）

- 将聚类结果与参考标准进行对比来评估聚类的性能。对于数据集$\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\}$，假设聚类标准为$\mathcal{P} = \{P_1, P_2, \dots, P_m\}$、聚类结果是$\mathcal{C} = \{C_1, C_2, \dots, C_k\}$

- 对于$\mathcal{D}$中任意两个不同的样本$\boldsymbol{x}^{(i)}$和$\boldsymbol{x}^{(j)}$，根据它们隶属于$\mathcal{C}$和$\mathcal{P}$的情况，定义四种关系：
  - SS：$\boldsymbol{x}^{(i)}$和$\boldsymbol{x}^{(j)}$在$\mathcal{C}$中属于相同簇，在$\mathcal{P}$中也属于相同簇；
  - SD：$\boldsymbol{x}^{(i)}$和$\boldsymbol{x}^{(j)}$在$\mathcal{C}$中属于相同簇，在$\mathcal{P}$中属于不同簇；
  - DS：$\boldsymbol{x}^{(i)}$和$\boldsymbol{x}^{(j)}$在$\mathcal{C}$中属于不同簇，在$\mathcal{P}$中属于相同簇；
  - DD：$\boldsymbol{x}^{(i)}$和$\boldsymbol{x}^{(j)}$在$\mathcal{C}$中属于不同簇，在$\mathcal{P}$中也属于不同簇。

- Rand统计量

$$RS = \frac{\#SS + \#DD}{\#SS + \#SD + \#DS + \#DD}$$

- Jaccard系数

$$JC = \frac{\#SS}{\#SS + \#SD + \#DS}$$

- FM指数（Fowlkers and Mallows index）

$$FMI = \sqrt{\frac{\#SS}{\#SS + \#SD} \cdot \frac{\#SS}{\#SS + \#DS}}$$

聚类性能评估（外部标准）

- 为了对聚类结果进行更加微观地评估，可以针对聚类标准的每一簇P_j和聚类结果的每一簇C_i，分别定义以下微观指标：
  - 精确率（precision）
  - 召回率（recall）
  - F_1值

$$P(P_j, C_i) = \frac{|P_j \cap C_i|}{|C_i|} \qquad R(P_j, C_i) = \frac{|P_j \cap C_i|}{|P_j|} \qquad F_1(P_j, C_i) = \frac{2 \cdot P(P_j, C_i) \cdot R(P_j, C_i)}{P(P_j, C_i) + R(P_j, C_i)}$$

- 对于聚类参考标准中的每个簇P_j

$$F_1(P_j) = \max_i \{F_1(P_j, C_i)\}$$

- 对于整个聚类

$$F_1 = \frac{\sum_j \left(|P_j| \cdot F_1(P_j)\right)}{\sum_j |P_j|}$$

聚类性能评估（内部标准）

- 仅靠考察聚类本身的分布结构来评估聚类的性能。主要思路：簇间越分离（即相似度越低）越好，簇内越凝聚（即相似度越高）越好。
- 凝聚度：计算样本$\boldsymbol{x}$与其所在簇C_m中其他样本的平均距离

$$a(\boldsymbol{x}) = \frac{\sum_{x \in C_m, x \neq x'} d(\boldsymbol{x}, \boldsymbol{x}')}{|C_m| - 1}$$

反映d所属簇的凝聚度，值越小越凝聚

- 分离度：计算样本$\boldsymbol{x}$与其它簇中样本的最小平均距离

$$b(\boldsymbol{x}) = \min_{C_j: 1 \leq j \leq k, j \neq m} \left\{ \frac{\sum_{x' \in C_j} d(\boldsymbol{x}, \boldsymbol{x}')}{|C_j|} \right\}$$

反映d与其他簇的分离度，值越大越分离

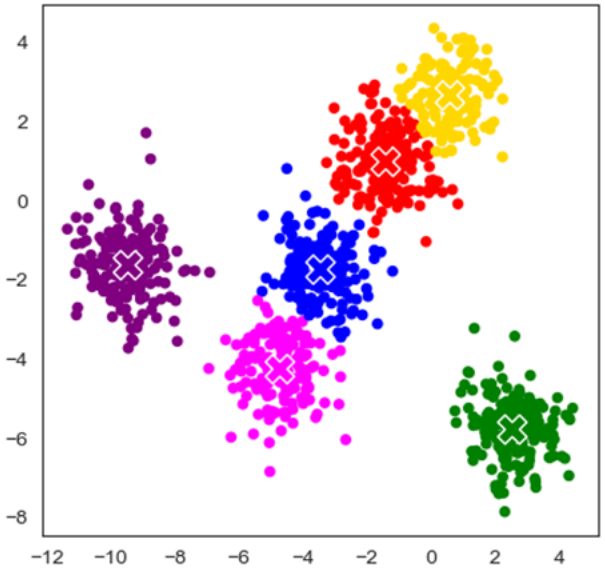
- 轮廓系数（Silhouette Coefficient）

$$SC(\boldsymbol{x}) = \frac{b(\boldsymbol{x}) - a(\boldsymbol{x})}{\max\{a(\boldsymbol{x}), b(\boldsymbol{x})\}} \Rightarrow SC = \frac{1}{N} \sum_{k=1}^{N} SC(\boldsymbol{x}^{(k)})$$

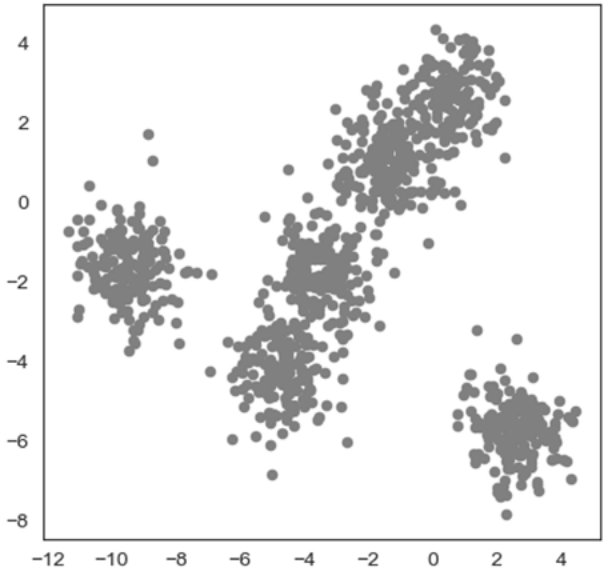
值域为$[-1,1]$，值越大说明聚类效果越好

作业#4

(1) GMM 分类数据集



(2) GMM 聚类数据集



http://www.nustm.cn/member/rxia/ml/data/gmm.zip

- 在GMM分类数据集上，实现基于高斯分布假设的贝叶斯模型，绘制分类曲线，报告分类正确率（5倍交叉验证）；
- 在GMM聚类数据集（去除GMM分类数据集的类别标签）上，实现：1) K-Means算法；2) GMM算法，绘制算法的动态聚类结果，将两种算法进行比较，并报告Rand统计量、FM指数、轮廓系数。

本讲结束 欢迎提问