

## 九、聚类

主讲教师：赵鹏

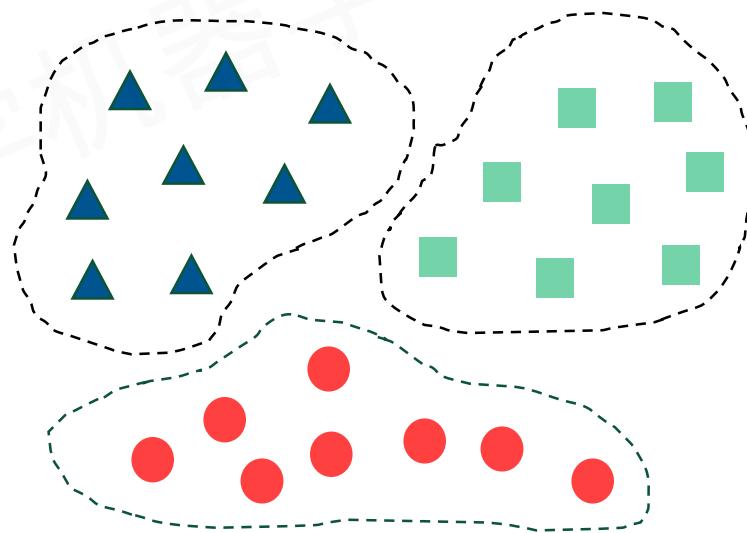
# 聚类 (Clustering)

在“无监督学习”任务中研究最多、应用最广

目标：将数据样本划分为若干个通常不相交的“簇”(cluster)

既可以作为一个单独过程（用于找寻数据内在的分布结构）

也可作为分类等其他学习任务的前驱过程



# 性能度量

聚类性能度量，亦称聚类“有效性指标”(validity index)

## □ 外部指标 (external index)

将聚类结果与某个“参考模型”(reference model)进行比较  
如 Jaccard 系数, FM 指数, Rand 指数

- Jaccard 系数(Jaccard Coefficient, 简称 JC)

$$JC = \frac{a}{a + b + c} . \quad (9.5)$$

- FM 指数(Fowlkes and Mallows Index, 简称 FMI)

$$FMI = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}} . \quad (9.6)$$

- Rand 指数(Rand Index, 简称 RI)

$$RI = \frac{2(a + d)}{m(m - 1)} . \quad (9.7)$$

显然, 上述性能度量的结果值均在 [0, 1] 区间, 值越大越好.

# 性能度量

聚类性能度量，亦称聚类“有效性指标”(validity index)

## □ 外部指标 (external index)

将聚类结果与某个“参考模型”(reference model)进行比较  
如 Jaccard 系数, FM 指数, Rand 指数

对数据集  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ , 假定通过聚类给出的簇划分为  $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ , 参考模型给出的簇划分为  $\mathcal{C}^* = \{C_1^*, C_2^*, \dots, C_s^*\}$ . 相应地, 令  $\lambda$  与

基本想法

没有“绝对正确”的标记/集簇信息, 相比外部聚类模型,  
使用“成对样本的聚类一致性”作为参考。

$$d = |DD|, \quad DD = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}, \quad (9.4)$$

# 性能度量

聚类性能度量，亦称聚类“有效性指标”(validity index)

## □ 内部指标 (internal index)

直接考察聚类结果而不用任何参考模型  
如 DB 指数, Dunn 指数等

基于式(9.8)~(9.11)可导出下面这些常用的聚类性能度量内部指标:

- DB 指数(Davies-Bouldin Index, 简称 DBI)

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left( \frac{\text{avg}(C_i) + \text{avg}(C_j)}{d_{\text{cen}}(C_i, C_j)} \right). \quad (9.12)$$

- Dunn 指数(Dunn Index, 简称 DI)

$$\text{DI} = \min_{1 \leqslant i \leqslant k} \left\{ \min_{j \neq i} \left( \frac{d_{\min}(C_i, C_j)}{\max_{1 \leqslant l \leqslant k} \text{diam}(C_l)} \right) \right\}. \quad (9.13)$$

显然, DBI 的值越小越好, 而 DI 则相反, 值越大越好.

# 性能度量

聚类性能度量，亦称聚类“有效性指标”(validity index)

## □ 内部指标 (internal index)

直接考察聚类结果而不用任何参考模型  
如 DB 指数, Dunn 指数等

考虑聚类结果的簇划分  $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ , 定义

$$\text{avg}(\mathcal{C}) = \frac{2}{|C|(|C| - 1)} \sum_{1 \leq i < j \leq |C|} \text{dist}(\mathbf{x}_i, \mathbf{x}_j), \quad (9.8)$$

### 基本想法

- “簇内相似度” (intra-cluster similarity) 高, 且
- “簇间相似度” (inter-cluster similarity) 低

$\overline{|C|} \sum_{1 \leq i \leq |C|} \mathbf{x}_i$ . 依然,  $\text{avg}(\mathcal{C})$  对应于簇  $\mathcal{C}$  内样本间的平均距离,  $d_{max}(\mathcal{C})$  对应于簇  $C$  内样本间的最远距离,  $d_{min}(C_i, C_j)$  对应于簇  $C_i$  与簇  $C_j$  最近样本间的距离,  $d_{cen}(C_i, C_j)$  对应于簇  $C_i$  与簇  $C_j$  中心点间的距离.

# 性能度量

---

聚类性能度量，亦称聚类“有效性指标”(validity index)

## □ 外部指标 (external index)

将聚类结果与某个“参考模型”(reference model)进行比较  
如 Jaccard 系数, FM 指数, Rand 指数

## □ 内部指标 (internal index)

直接考察聚类结果而不用任何参考模型  
如 DB 指数, Dunn 指数等



上述的讨论基础：一个合适“**距离度量**”

# 距离计算

距离度量 (distance metric) 需满足的基本性质：

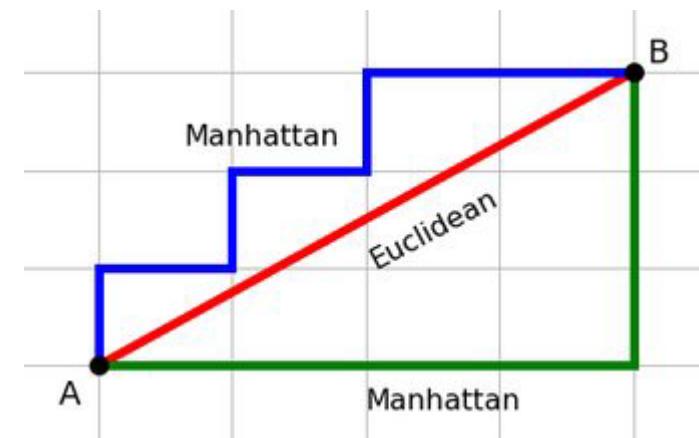
- 非负性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ ;
- 同一性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = 0$  当且仅当  $\mathbf{x}_i = \mathbf{x}_j$ ;
- 对称性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = \text{dist}(\mathbf{x}_j, \mathbf{x}_i)$ ;
- 直递性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \text{dist}(\mathbf{x}_i, \mathbf{x}_k) + \text{dist}(\mathbf{x}_k, \mathbf{x}_j)$ .

常用距离形式：

闵可夫斯基距离 (Minkowski distance)

$$\text{dist}_{mk}(\mathbf{x}_i, \mathbf{x}_j) = \left( \sum_{u=1}^n |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

- $p = 2$ : 欧氏距离(Euclidean distance)
- $p = 1$ : 曼哈顿距离(Manhattan distance)

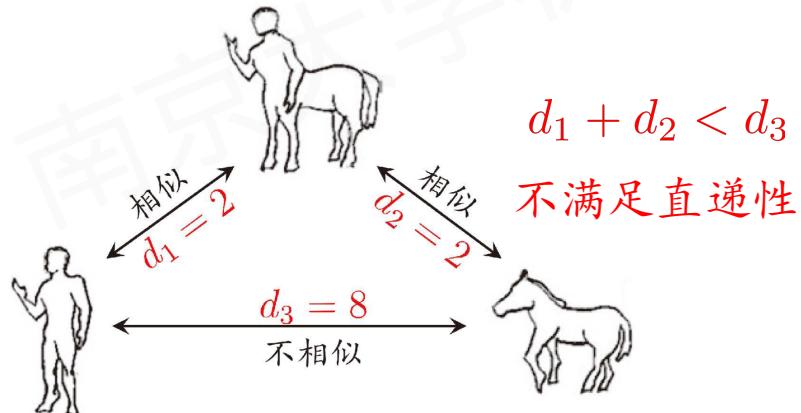


# 距离计算

距离度量 (distance metric) 需满足的基本性质：

- 非负性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ ;
- 同一性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = 0$  当且仅当  $\mathbf{x}_i = \mathbf{x}_j$ ;
- 对称性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = \text{dist}(\mathbf{x}_j, \mathbf{x}_i)$ ;
- 直递性:  $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \text{dist}(\mathbf{x}_i, \mathbf{x}_k) + \text{dist}(\mathbf{x}_k, \mathbf{x}_j)$ .

“非度量” 距离 (non-metric distance)



$$\text{dist}_{\text{sqE}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

$$\|-2 - 2\|_2^2 \cancel{\leq} \| -2 - 0 \|_2^2 + \| 0 - 2 \|_2^2$$

# 距离计算

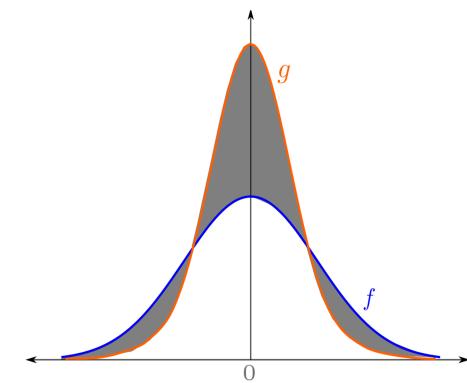
- 对无序(non-ordinal)属性，可使用 VDM (Value Difference Metric)

令  $m_{u,a}$  表示属性  $u$  上取值为  $a$  的样本数， $m_{u,a,i}$  表示在第  $i$  个样本簇中在属性  $u$  上取值为  $a$  的样本数， $k$  为样本簇数，则属性  $u$  上两个离散值  $a$  与  $b$  之间的 VDM 距离为

$$\text{VDM}_p(a, b) = \sum_{i=1}^k \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p$$

可认为是对两个离散属性值  $a$  与  $b$  进行编码表示

- 离散属性值  $a$  的编码表示： $\left[ \frac{m_{u,a,1}}{m_{u,a}}, \frac{m_{u,a,2}}{m_{u,a}}, \dots, \frac{m_{u,a,k}}{m_{u,a}} \right]$
- 离散属性值  $b$  的编码表示： $\left[ \frac{m_{u,b,1}}{m_{u,b}}, \frac{m_{u,b,2}}{m_{u,b}}, \dots, \frac{m_{u,b,k}}{m_{u,b}} \right]$



# 距离计算

- 对无序(non-ordinal)属性，可使用 VDM (Value Difference Metric)

令  $m_{u,a}$  表示属性  $u$  上取值为  $a$  的样本数， $m_{u,a,i}$  表示在第  $i$  个样本簇中在属性  $u$  上取值为  $a$  的样本数， $k$  为样本簇数，则属性  $u$  上两个离散值  $a$  与  $b$  之间的 VDM 距离为

$$\text{VDM}_p(a, b) = \sum_{i=1}^k \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p$$

- 对混合属性，可使用 MinkovDM

$$\text{MinkovDM}_p(x_i, x_j) = \left( \sum_{u=1}^{n_c} |x_{iu} - x_{ju}|^p + \sum_{u=n_c+1}^n \text{VDM}_p(x_{iu}, x_{ju}) \right)^{\frac{1}{p}}$$

# 必须记住



聚类的“好坏”不存在绝对标准

**the goodness of clustering depends on  
the opinion of the user**

# 故事一则 (from 周老师)

---

聚类的故事：

老师拿来苹果和梨，让小朋友分成两份。

小明把大苹果大梨放一起，小个头的放一起，老师点头，恩，  
体量感。

小芳把红苹果挑出来，剩下的放一起，老师点头，颜色感。

小武的结果？不明白。小武掏出眼镜：最新款，能看到水果里  
有几个籽，左边这堆单数，右边双数。

老师很高兴：新的聚类算法诞生了

聚类也许是机器学习中“新算法”出现最多、最快的领域  
总能找到一个新的“标准”，使以往算法对它无能为力

# 常见聚类方法

---

## □ 原型聚类

- 亦称“基于原型的聚类”(prototype-based clustering)
- 假设：聚类结构能通过一组原型刻画
- 过程：先对原型初始化，然后对原型进行迭代更新求解
- 代表：**k均值聚类，学习向量量化(LVQ)，高斯混合聚类**

## □ 密度聚类

- 亦称“基于密度的聚类”(density-based clustering)
- 假设：聚类结构能通过样本分布的紧密程度确定
- 过程：从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇
- 代表：**DBSCAN, OPTICS, DENCLUE**

## □ 层次聚类 (hierarchical clustering)

- 假设：能够产生不同粒度的聚类结果
- 过程：在不同层次对数据集进行划分，从而形成树形的聚类结构
- 代表：**AGNES (自底向上), DIANA (自顶向下)**

# 原型聚类：k-means

每个簇以该簇中所有样本点的“均值”表示

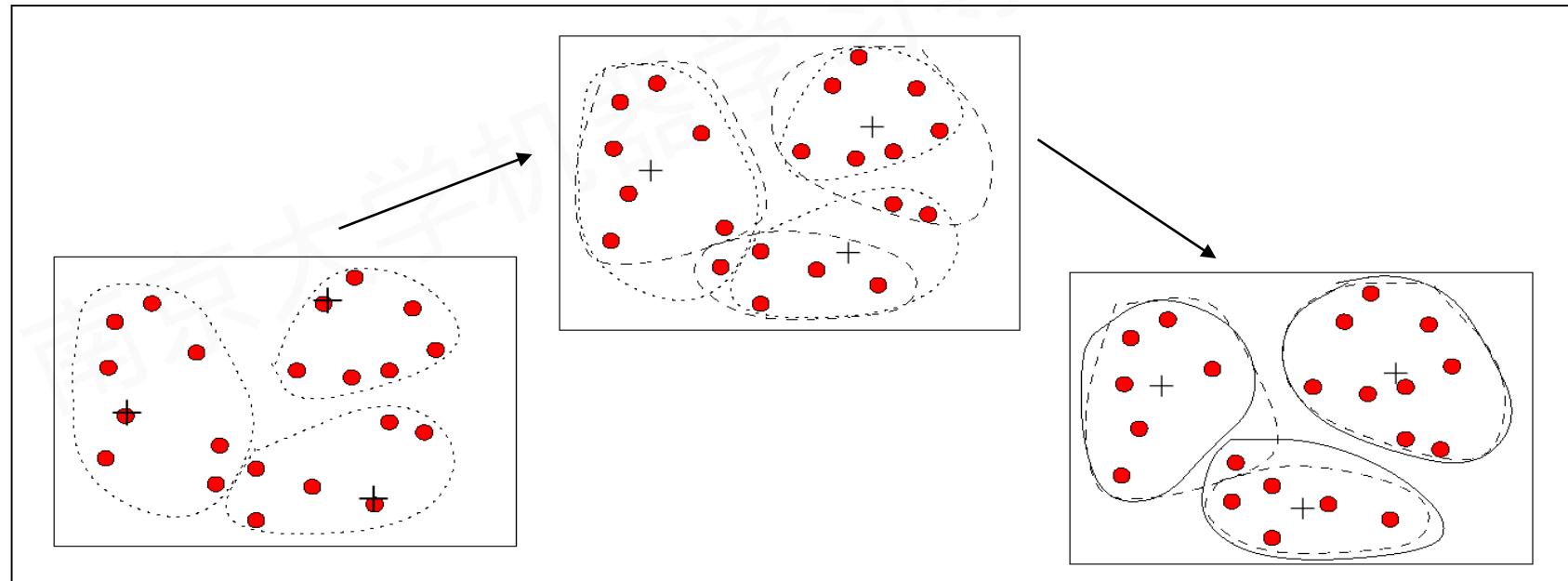
若不以均值向量为原型，而是以距离它最近的样本点为原型，则得到 k-medoids 算法

Step1: 随机选取k个样本点作为簇中心

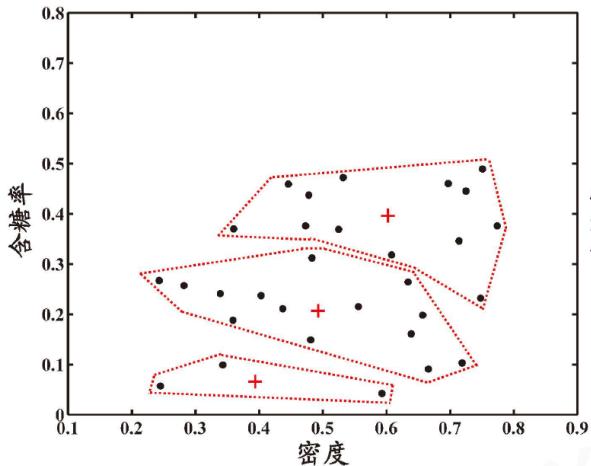
Step2: 将其他样本点根据其与簇中心的距离，划分给最近的簇

Step3: 更新各簇的均值向量，将其作为新的簇中心

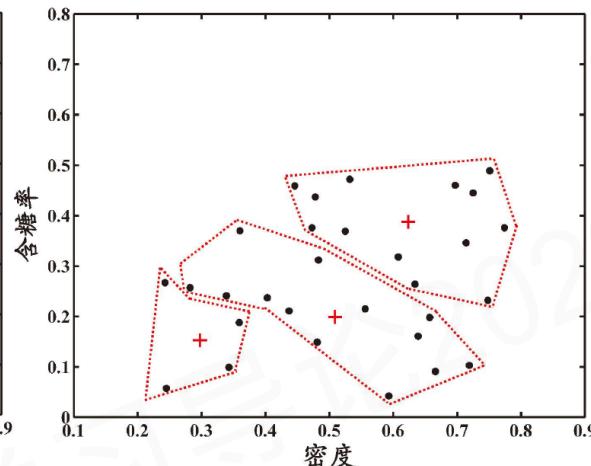
Step4: 若所有簇中心未发生改变，则停止；否则执行 Step 2



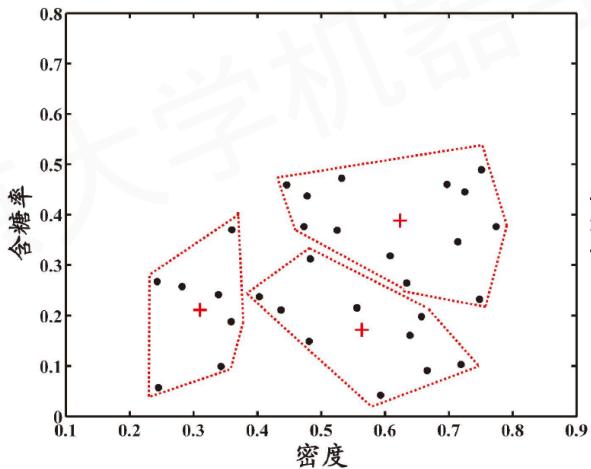
# 原型聚类：k-means



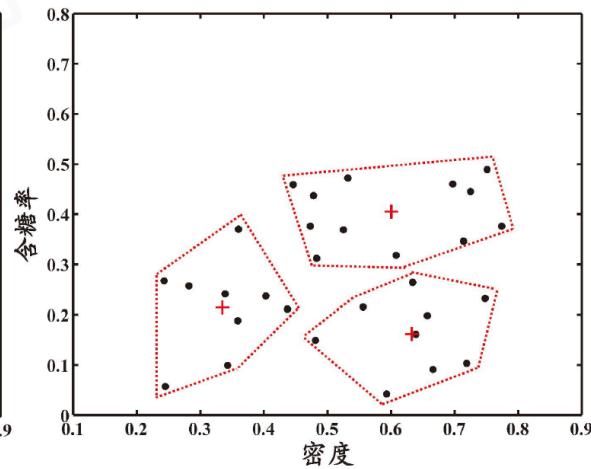
(a) 第一轮迭代后



(b) 第二轮迭代后



(c) 第三轮迭代后



(d) 第四轮迭代后

# 原型聚类：学习向量量化 (Learning Vector Quantization, LVQ)

也是试图找到一组原型向量来刻画聚类结构，但假设数据样本带有类别标记  
实际上是通过聚类来形成类别的“子类”结构，每个子类对应一个聚类簇

---

**输入:** 样本集  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ ;  
原型向量个数  $q$ , 各原型向量预设的类别标记  $\{t_1, t_2, \dots, t_q\}$ ;  
学习率  $\eta \in (0, 1)$ .

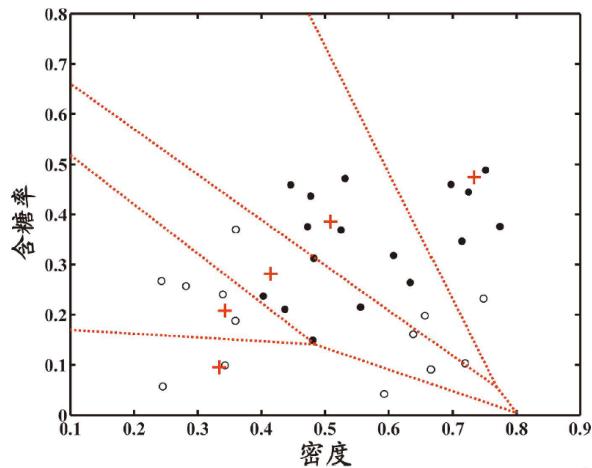
**过程:**

- 1: 初始化一组原型向量  $\{p_1, p_2, \dots, p_q\}$
- 2: **repeat**
- 3:   从样本集  $D$  随机选取样本  $(x_j, y_j)$ ;
- 4:   计算样本  $x_j$  与  $p_i$  ( $1 \leq i \leq q$ ) 的距离:  $d_{ji} = \|x_j - p_i\|_2$ ;
- 5:   找出与  $x_j$  距离最近的原型向量  $p_{i^*}$ ,  $i^* = \arg \min_{i \in \{1, 2, \dots, q\}} d_{ji}$ ;
- 6:   **if**  $y_j = t_{i^*}$  **then**
- 7:      $p' = p_{i^*} + \eta \cdot (x_j - p_{i^*})$     $x_j$  与  $p_{i^*}$  的类别相同
- 8:   **else**
- 9:      $p' = p_{i^*} - \eta \cdot (x_j - p_{i^*})$     $x_j$  与  $p_{i^*}$  的类别不同
- 10:   **end if**
- 11:   将原型向量  $p_{i^*}$  更新为  $p'$
- 12: **until** 满足停止条件

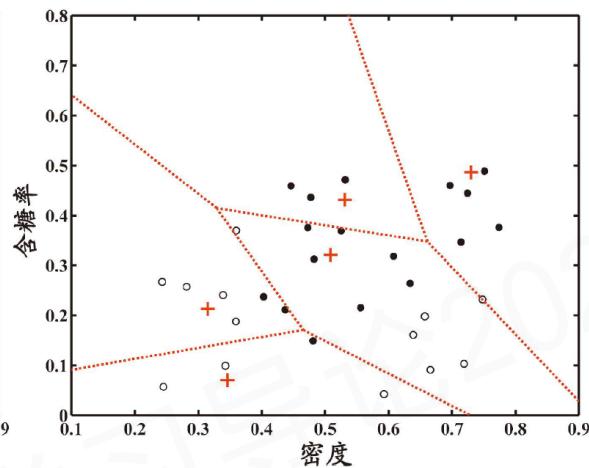
**输出:** 原型向量  $\{p_1, p_2, \dots, p_q\}$

---

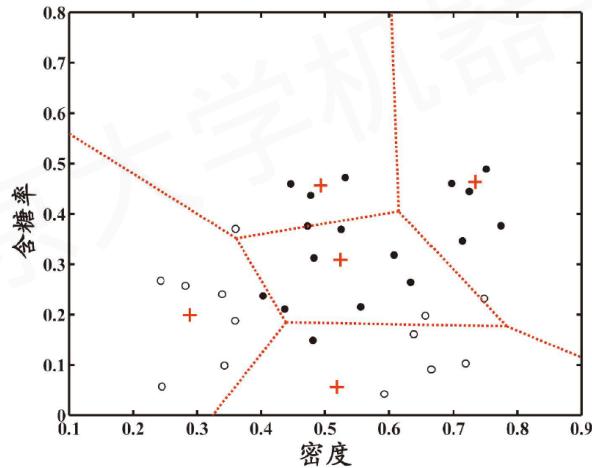
# 原型聚类：学习向量量化 (Learning Vector Quantization, LVQ)



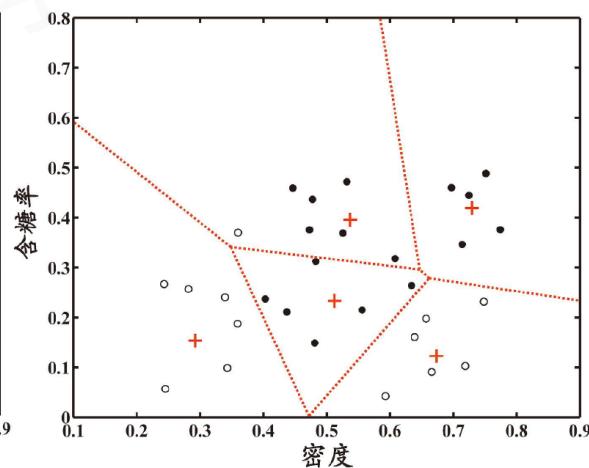
(a) 50 轮迭代后



(b) 100 轮迭代后



(c) 200 轮迭代后



(d) 400 轮迭代后

# 原型聚类：高斯混合聚类 (Gaussian Mixture Clustering, GMM)

采用概率模型来表达聚类原型

$n$  维样本空间中的随机向量  $\mathbf{x}$  若服从高斯分布，则其概率密度函数为

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

假设样本由下面这个高斯混合分布生成：

生成式模型

$$p_{\mathcal{M}}(\mathbf{x}) = \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x} | \boldsymbol{\mu}_i, \Sigma_i)$$

- 根据  $\alpha_1, \alpha_2, \dots, \alpha_k$  定义的先验分布选择高斯混合成分，其中  $\alpha_i$  为选择第  $i$  个混合成分的概率；
- 然后，根据被选择的混合成分的概率密度函数进行采样，从而生成相应的样本

## 补充：EM算法

如何处理“未观测到的”变量？

例如，西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”，  
则训练样本的“根蒂”属性变量值未知

未观测变量 → 隐变量 (latent variable)

EM (Expectation-Maximization) 算法是估计隐变量的利器

令  $\mathbf{X}$  表示已观测变量集， $\mathbf{Z}$  表示隐变量集，欲对模型参数  $\Theta$  做极大似然估计，则应最大化对数似然函数

$$LL(\Theta \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \Theta)$$

$\mathbf{Z}$  是隐变量，无法直接求解。怎么办？

## 补充：EM算法

对隐变量  $\mathbf{Z}$  计算期望，根据训练数据最大化对数“边际似然”  
(marginal likelihood)

$$LL(\Theta | \mathbf{X}) = \ln P(\mathbf{X} | \Theta) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} | \Theta)$$

以初始值  $\Theta^0$  为起点，迭代执行以下步骤直至收敛：

- 基于  $\Theta^t$  推断隐变量  $\mathbf{Z}$  的期望，记为  $\mathbf{Z}^t$
- 基于已观测变量  $\mathbf{X}$  和  $\mathbf{Z}^t$  对参数  $\Theta$  做极大似然估计，记为  $\Theta^{t+1}$

**E步**: 当  $\Theta$  已知  $\rightarrow$  根据训练数据推断隐变量  $\mathbf{Z}$

**M步**: 当  $\mathbf{Z}$  已知  $\rightarrow$  对  $\Theta$  做极大似然估计

一般形式：**E-M** 两个步骤交替计算，直至收敛：

- **E步 - 计算期望**: 利用当前估计的参数值计算对数似然的期望值；
- **M步 - 最大化**: 寻找能使**E步**产生的似然期望最大化的参数值；

## 高斯混合聚类：EM求解

引入**隐变量**  $z_j \in [k]$  ( $j = 1, 2, \dots, m$ ) 表示生成样本  $x_j$  的高斯混合成分

- 已观测变量集  $\mathbf{X} = \{x_j\}_{j \in [m]}$
- 隐变量集  $\mathbf{Z} = \{z_j\}_{j \in [m]}$
- 模型参数  $\Theta = \{(\alpha_i, \mu_i, \Sigma_i)\}_{i \in [k]}$

**E步**: 当  $\Theta$  已知  $\rightarrow$  根据训练数据推断隐变量  $\mathbf{Z}$

**M步**: 当  $\mathbf{Z}$  已知  $\rightarrow$  对  $\Theta$  做极大似然估计

## 高斯混合聚类：EM求解

引入**隐变量**  $z_j \in [k]$  ( $j = 1, 2, \dots, m$ ) 表示生成样本  $\mathbf{x}_j$  的高斯混合成分

**E步**: 当  $\Theta$  已知  $\rightarrow$  根据训练数据推断**隐变量 Z**

样本  $\mathbf{x}_j$  由第  $i$  个高斯混合成分生成的后验概率为：

$$p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) = \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\mathbf{x}_j \mid z_j = i)}{p_{\mathcal{M}}(\mathbf{x}_j)} = \frac{\alpha_i \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

简记为  $\gamma_{ji} \triangleq p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j)$  ( $i = 1, 2, \dots, k$ )

## 高斯混合聚类：EM求解

引入**隐变量**  $z_j \in [k]$  ( $j = 1, 2, \dots, m$ ) 表示生成样本  $\mathbf{x}_j$  的高斯混合成分

**M步：**当  $\mathbf{Z}$  已知  $\rightarrow$  对  $\Theta$  做极大似然估计

极大似然法，由高斯混合假设，可得到对数似然，进行最大化

$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

# 高斯混合聚类：EM求解

- 样本  $\mathbf{x}_j$  由第  $i$  个高斯混合成分生成的后验概率为：

$$p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) = \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\mathbf{x}_j \mid z_j = i)}{p_{\mathcal{M}}(\mathbf{x}_j)} = \frac{\alpha_i \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

简记为  $\gamma_{ji} \triangleq p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j)$  ( $i = 1, 2, \dots, k$ )

- 极大似然法，由高斯混合假设，可得到对数似然，进行最大化：

$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

**EM 算法：**

- (E步) 根据当前参数计算每个样本属于每个高斯成分的后验概率  $\gamma_{ji}$
- (M步) 更新模型参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$

## 高斯混合聚类：M步求解

- 极大似然法，由高斯混合假设，可得到对数似然，进行最大化：

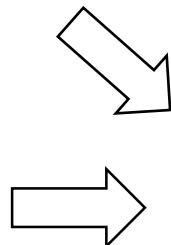
$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

若参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$  能使得上式最大化，则由  $\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = 0$  可得

$$\sum_{j=1}^m \frac{\alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{x}_j - \boldsymbol{\mu}_i) = 0$$

E步得到每个成分后验概率

$$\begin{aligned}\gamma_{ji} &\triangleq p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) \\ &= \frac{\alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}\end{aligned}$$



$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^m \gamma_{ji} \mathbf{x}_j}{\sum_{j=1}^m \gamma_{ji}}$$

## 高斯混合聚类：M步求解

- 极大似然法，由高斯混合假设，可得到对数似然，进行最大化：

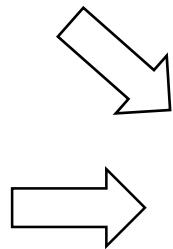
$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

若参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$  能使得上式最大化，则由  $\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = 0$  可得

$$\sum_{j=1}^m \frac{\alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{x}_j - \boldsymbol{\mu}_i) = 0$$

E步得到每个成分后验概率

$$\begin{aligned}\gamma_{ji} &\triangleq p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) \\ &= \frac{\alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}\end{aligned}$$


$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^m \gamma_{ji} (\mathbf{x}_j - \boldsymbol{\mu}_i) (\mathbf{x}_j - \boldsymbol{\mu}_i)^{\top}}{\sum_{j=1}^m \gamma_{ji}}$$

## 高斯混合聚类：M步求解

- 极大似然法，由高斯混合假设，可得到对数似然，进行最大化：

$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

对于参数 $\{\alpha_i\}_{i \in [k]}$ ，还需满足 $\alpha_i \geq 0, \sum_{i=1}^k \alpha_i = 1$ ，可采用拉格朗日乘子法

$LL(D)$  的拉格朗日形式

$$LL(D) + \lambda \left( \sum_{i=1}^k \alpha_i - 1 \right), \quad (9.36)$$

其中 $\lambda$  为拉格朗日乘子。由式(9.36)对 $\alpha_i$  的导数为 0，有

$$\sum_{j=1}^m \frac{p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda = 0, \quad (9.37)$$

两边同乘以 $\alpha_i$ ，对所有混合成分求和可知 $\lambda = -m$ ，有

$$\rightarrow \alpha_i = \frac{1}{m} \sum_{j=1}^m \gamma_{ji}$$

## 高斯混合聚类：M步求解

- 极大似然法，由高斯混合假设，可得到对数似然，进行最大化：

$$LL(D) = \ln \left( \prod_{j=1}^m p_{\mathcal{M}}(\mathbf{x}_j) \right) = \sum_{j=1}^m \ln \left( \sum_{i=1}^k \alpha_i \cdot p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

若参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$  能使得上式最大化，最终综合得到

E步得到每个成分后验概率

$$\begin{aligned}\gamma_{ji} &\triangleq p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) \\ &= \frac{\alpha_i \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l \cdot p(\mathbf{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}\end{aligned}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^m \gamma_{ji} \mathbf{x}_j}{\sum_{j=1}^m \gamma_{ji}}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^m \gamma_{ji} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^{\top}}{\sum_{j=1}^m \gamma_{ji}}$$

$$\alpha_i = \frac{1}{m} \sum_{j=1}^m \gamma_{ji}$$

# 高斯混合聚类：最终算法

输入：样本集  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ ;  
高斯混合成分个数  $k$ .

过程：

- 1: 初始化高斯混合分布的模型参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$
- 2: **repeat**

EM 算法的 E 步.

- 3:   **for**  $j = 1, 2, \dots, m$  **do**
- 4:     根据式(9.30)计算  $\mathbf{x}_j$  由各混合成分生成的后验概率, 即  
 $\gamma_{ji} = p_M(z_j = i \mid \mathbf{x}_j)$  ( $1 \leq i \leq k$ )
- 5:   **end for**

EM 算法的 M 步.

- 6:   **for**  $i = 1, 2, \dots, k$  **do**
- 7:     计算新均值向量:  $\boldsymbol{\mu}'_i = \frac{\sum_{j=1}^m \gamma_{ji} \mathbf{x}_j}{\sum_{j=1}^m \gamma_{ji}}$ ;
- 8:     计算新协方差矩阵:  $\boldsymbol{\Sigma}'_i = \frac{\sum_{j=1}^m \gamma_{ji} (\mathbf{x}_j - \boldsymbol{\mu}'_i)(\mathbf{x}_j - \boldsymbol{\mu}'_i)^T}{\sum_{j=1}^m \gamma_{ji}}$ ;
- 9:     计算新混合系数:  $\alpha'_i = \frac{\sum_{j=1}^m \gamma_{ji}}{m}$ ;
- 10:   **end for**

例如达到最大迭代轮数.

- 11:   将模型参数  $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$  更新为  $\{(\alpha'_i, \boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i) \mid 1 \leq i \leq k\}$
- 12:   **until** 满足停止条件
- 13:    $C_i = \emptyset$  ( $1 \leq i \leq k$ )
- 14:   **for**  $j = 1, 2, \dots, m$  **do**
- 15:     根据式(9.31)确定  $\mathbf{x}_j$  的簇标记  $\lambda_j$ ;
- 16:     将  $\mathbf{x}_j$  划入相应的簇:  $C_{\lambda_j} = C_{\lambda_j} \cup \{\mathbf{x}_j\}$
- 17:   **end for**

输出：簇划分  $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$

# 高斯混合聚类

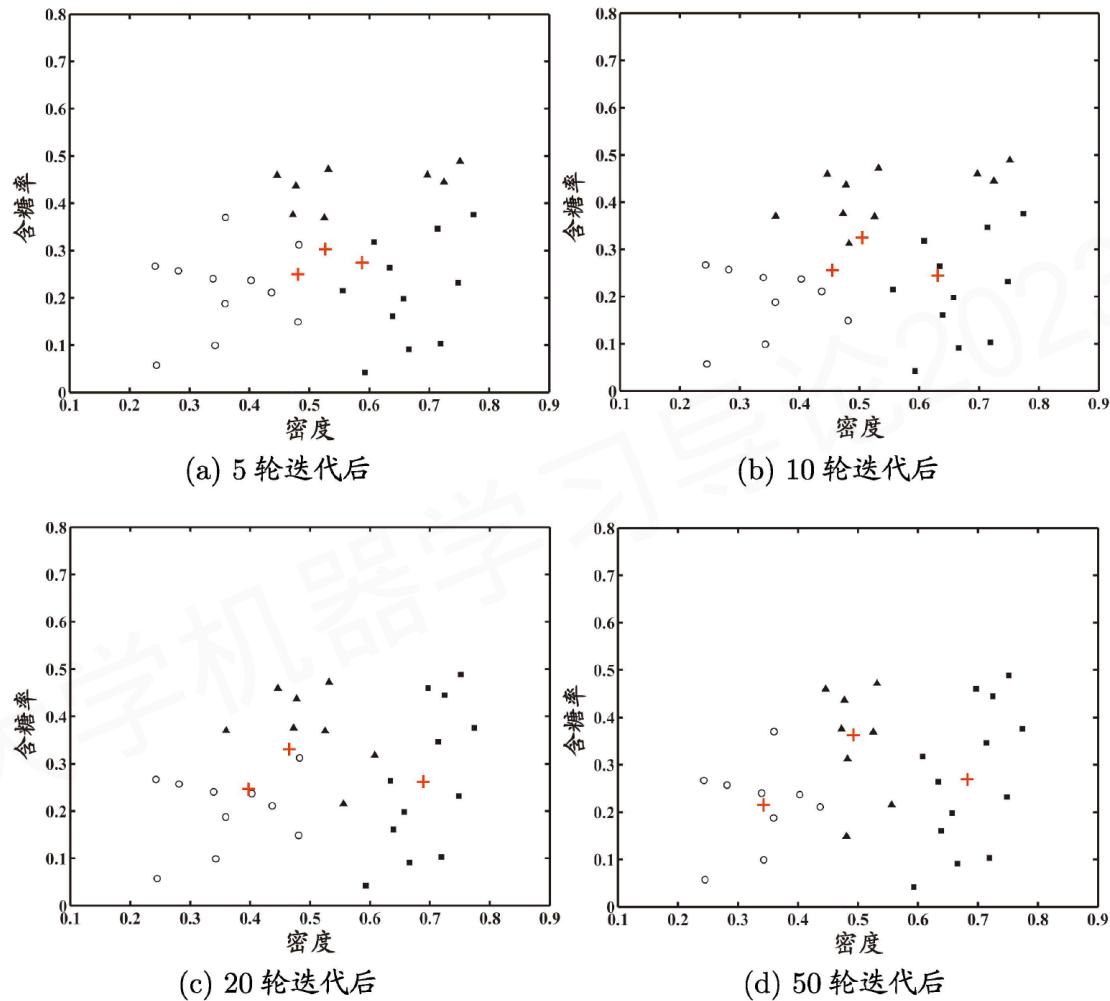


图 9.7 高斯混合聚类( $k = 3$ )在不同轮数迭代后的聚类结果. 其中样本簇  $C_1$ ,  $C_2$  与  $C_3$  中的样本点分别用“○”, “■”与“▲”表示, 各高斯混合成分的均值向量用“+”表示.

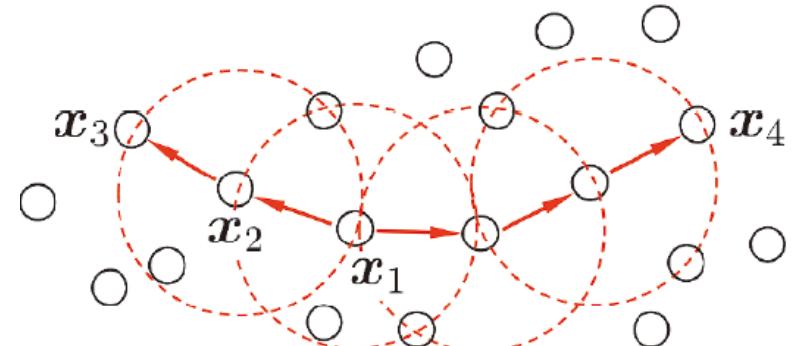
# 密度聚类：DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

关键概念：

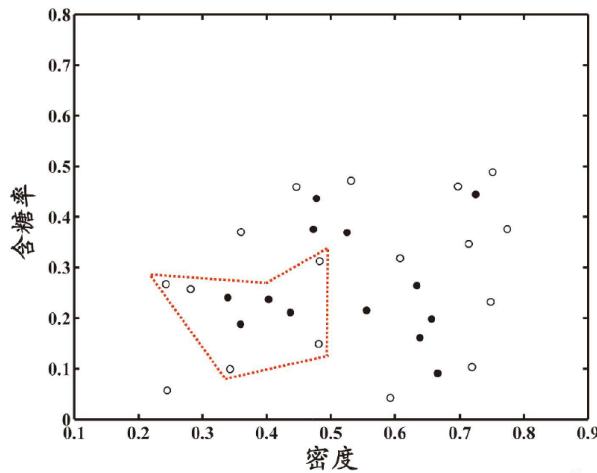
- 核心对象(core object): 若  $x_j$  的  $\epsilon$ -邻域至少包含  $MinPts$  个样本, 即  $|N_\epsilon(x_j)| \geq MinPts$ , 则  $x_j$  是一个核心对象;
- 密度直达(directly density-reachable): 若  $x_j$  位于  $x_i$  的  $\epsilon$ -邻域中, 且  $x_i$  是核心对象, 则称  $x_j$  由  $x_i$  密度直达;
- 密度可达(density-reachable): 对  $x_i$  与  $x_j$ , 若存在样本序列  $p_1, p_2, \dots, p_n$ , 其中  $p_1 = x_i$ ,  $p_n = x_j$  且  $p_{i+1}$  由  $p_i$  密度直达, 则称  $x_j$  由  $x_i$  密度可达;
- 密度相连(density-connected): 对  $x_i$  与  $x_j$ , 若存在  $x_k$  使得  $x_i$  与  $x_j$  均由  $x_k$  密度可达, 则称  $x_i$  与  $x_j$  密度相连.

令  $MinPts = 3$ ,  
虚线显示出  $\epsilon$  邻域  
 $x_1$  是核心对象

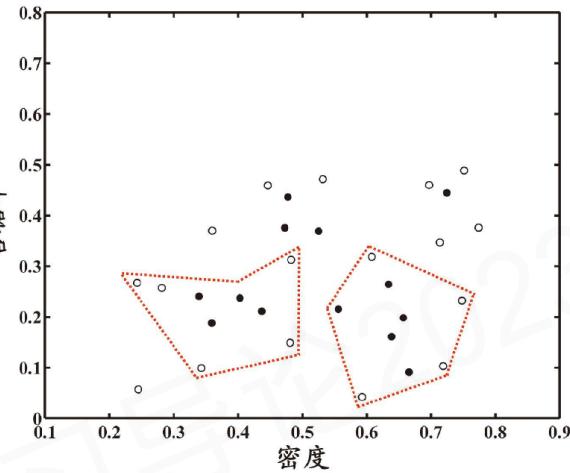
$x_2$  由  $x_1$  密度直达  
 $x_3$  由  $x_1$  密度可达  
 $x_3$  与  $x_4$  密度相连



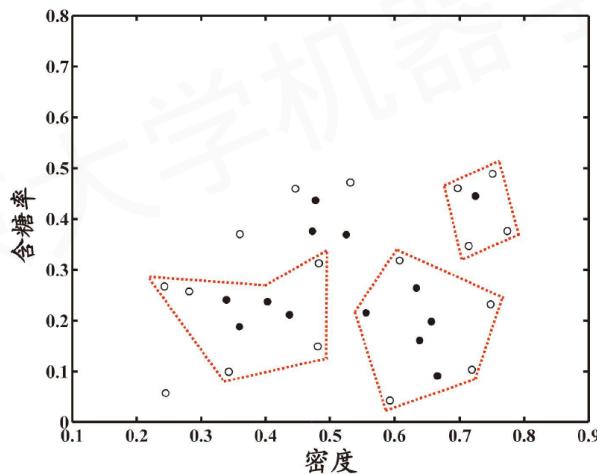
# 密度聚类：DBSCAN (Density-Based Spatial Clustering of Applications with Noise)



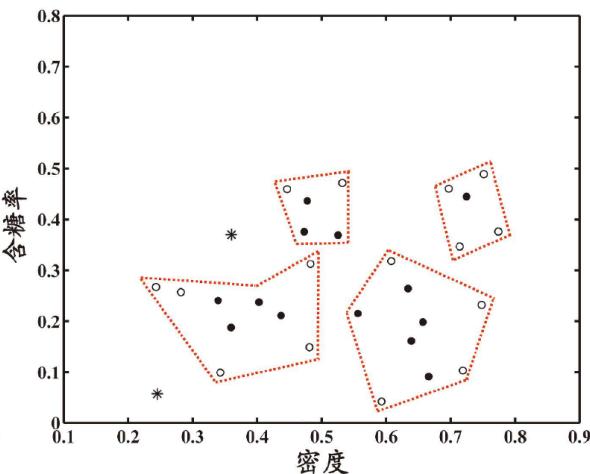
(a) 生成聚类簇  $C_1$



(b) 生成聚类簇  $C_2$



(c) 生成聚类簇  $C_3$



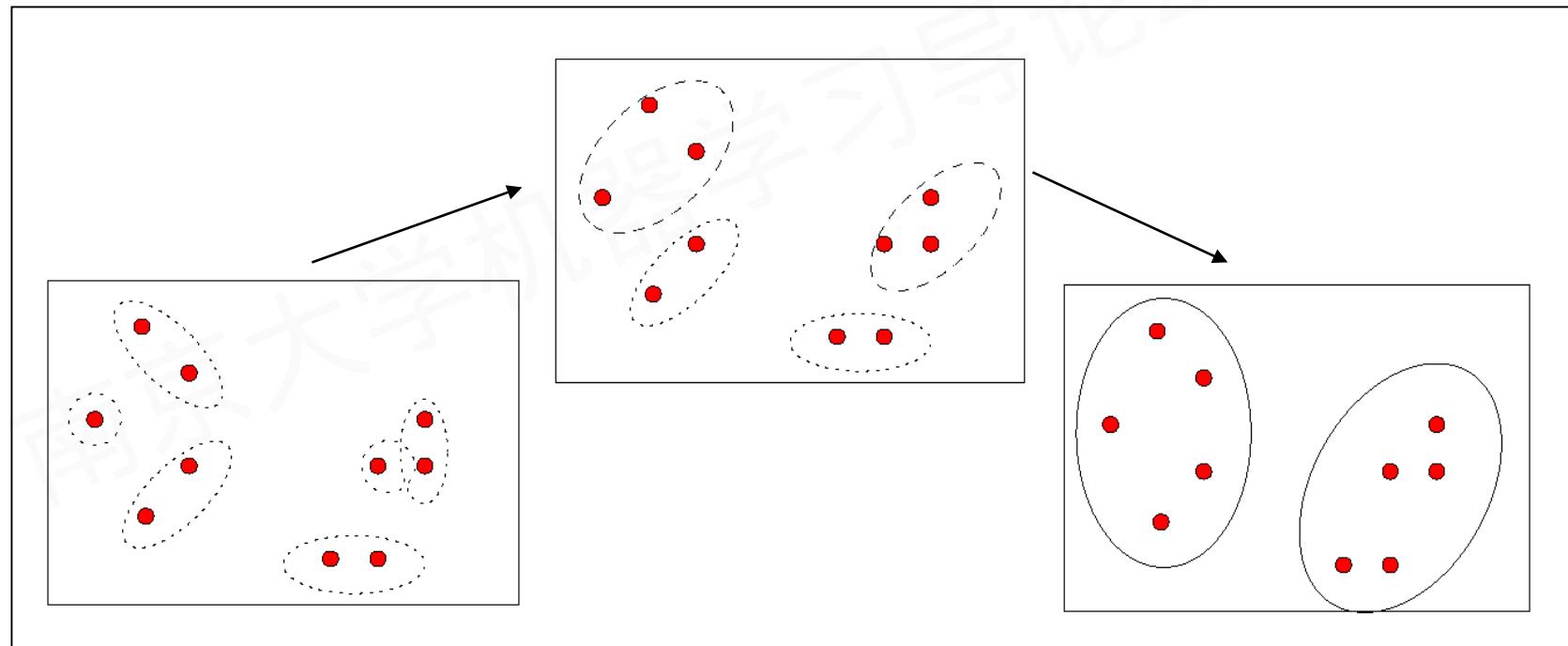
(d) 生成聚类簇  $C_4$

# 层次聚类：AGNES (AGglomerative NESting)

Step1: 将每个样本点作为一个簇

Step2: 合并最近的两个簇

Step3: 若所有样本点都存在与一个簇中，则停止；否则转到 Step2



# 层次聚类：AGNES (AGglomerative NESting)

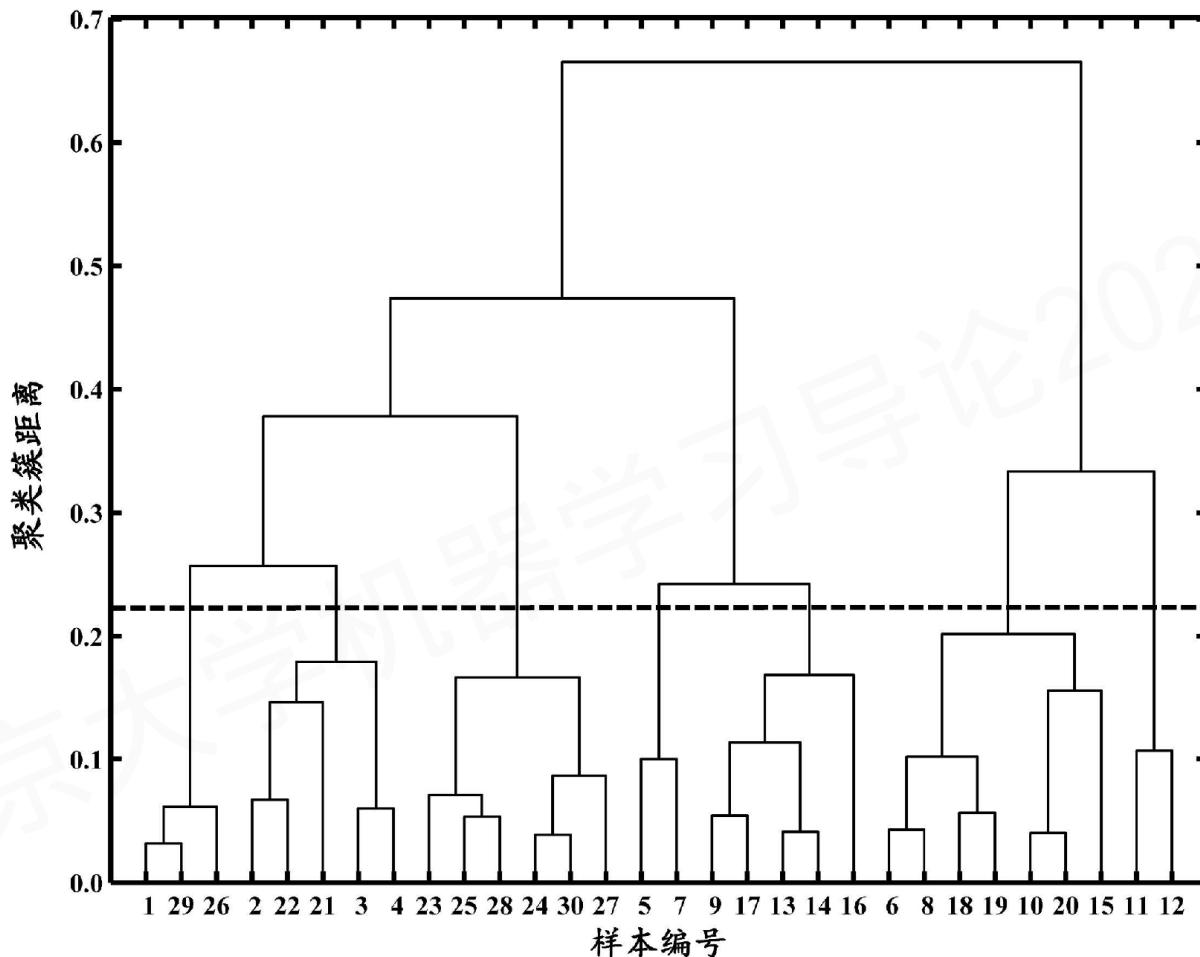


图 9.12 西瓜数据集 4.0 上 AGNES 算法生成的树状图(采用  $d_{max}$ ). 横轴对应于样本编号, 纵轴对应于聚类簇距离.

前往下一站.....

