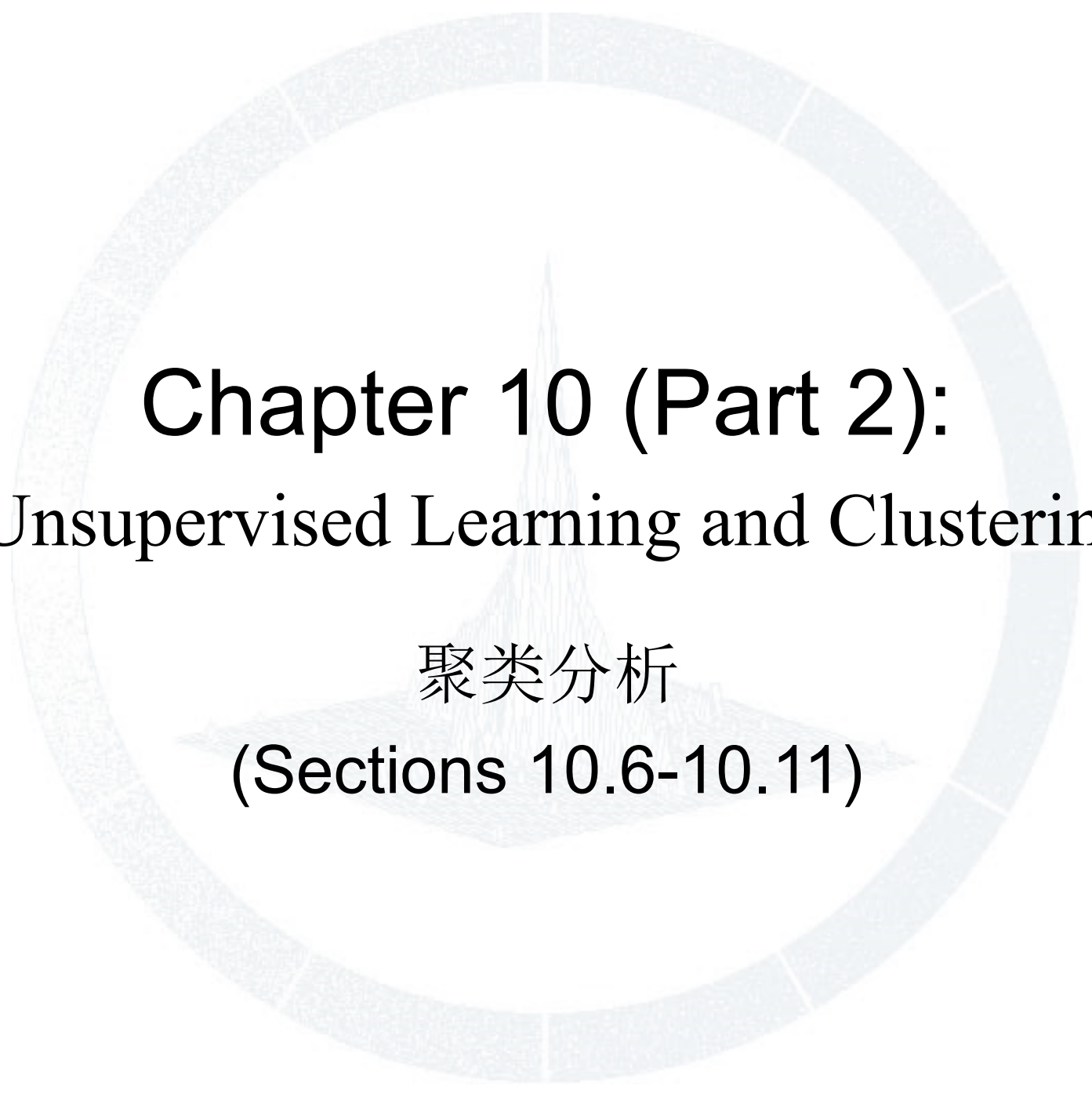




# 模式识别的理论与方法

## Pattern Recognition

裴继红



# Chapter 10 (Part 2): Unsupervised Learning and Clustering

聚类分析  
(Sections 10.6-10.11)

# 聚类分析

- **聚类，也称为无监督的分类**：对无标记样本集，定义样本之间的相似度，将相似样本的归为同一类。它是一种重要的人类行为，是模式分类和系统建模的基本方法之一。
  - **类（cluster）**：类是一组数据对象的集合，在同一个类中的对象彼此相似，而不同类中的对象彼此相异。
  - **聚类（clustering）**：将对象按照相似性分组为多个类的过程。
  - **聚类分析（clustering analysis）**：通过聚类对数据进行概括，或找到数据集中存在的自然的、真实的子结构。



# 数据描述与聚类

## ➤ 数据描述

- 对原始数据的统计假设，及由此计算出的统计量有时并不足以揭示数据集合的空间结构。
- 相同的低阶（如二阶）统计量可能对应多种不同的空间结构。
- 采用混合密度参数估计的方法在先验信息不足时会，可能导致对数据结构描述的错误的结果

## ➤ 聚类

- 聚类，是一种直接从数据中发现目标子类的方法。
- 每个子类中的样本具有高度的相似性。

## ➤ 基于目标函数优化的聚类方法

- 通过定义一个目标函数，并进行迭代优化，得到数据集的子类。



# 基于目标函数优化聚类的基本过程

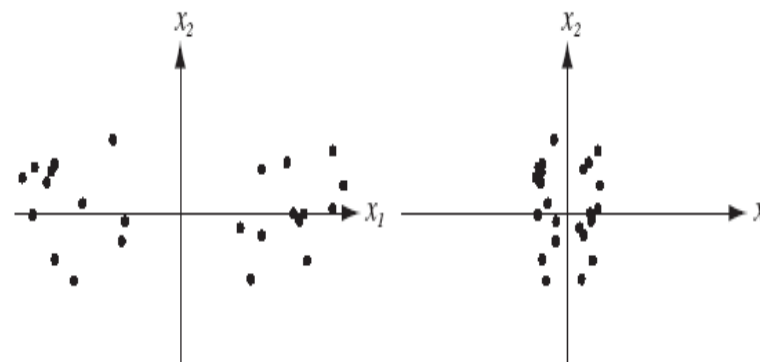
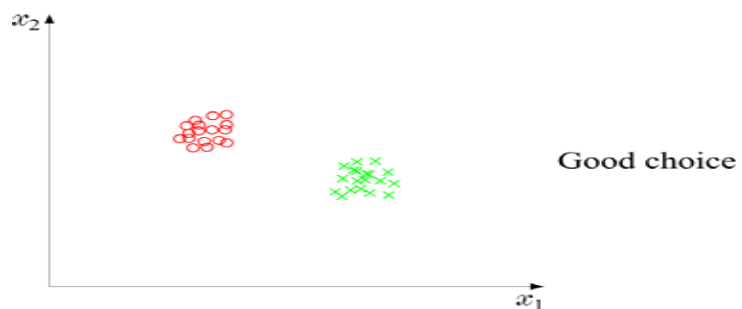
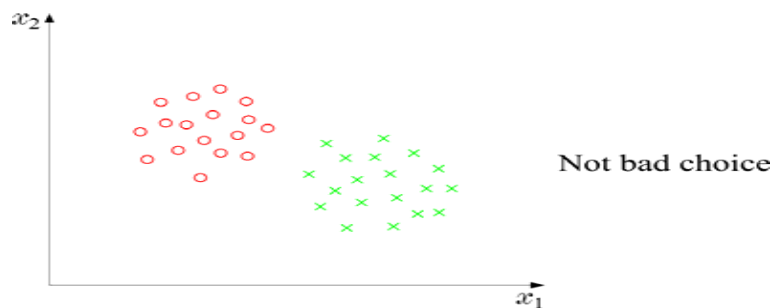
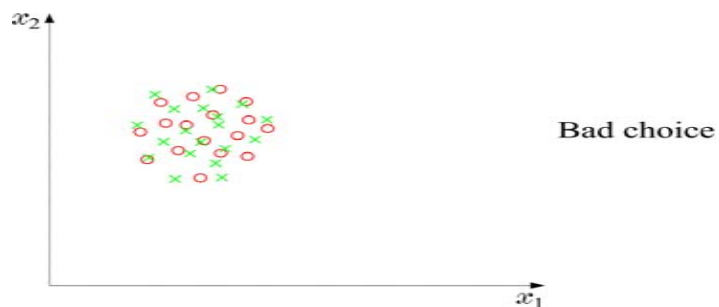
1. **特征选择**。选择特征尽可能与感兴趣的任務相关。特征之间的信息冗余度要尽可能小。
2. **相似性测度**。要求所有选择的特征对相似性测度计算的贡献都是均衡的，没有那一个特征是绝对占优的。
3. **聚类准则与准则函数**。聚类准则可以被表示为代价函数和其它类型的规则。它依赖于对数据集合内部隐含类的类型的判断与解释。
4. **聚类的优化算法**。在确定相似性测度和聚类准则后，选择一个具体的算法方案将数据集合分解为类结构。
5. **聚类结构的有效性**。在聚类算法获得了结果后，需要采用合适的检验方法检验其正确性。
6. **结果的解释**。应用领域的专家结合试验证据，分析解释聚类结果，以便得到正确的结论。



# 聚类：特征选择

- 特征选择：去除那些信息内容贫乏的单个特征，保留信息丰富的那些特征，并将它们组成一个向量。
  - 选择“最优”的特征数  $l$ 。
    - 大的  $l$  有三个方面的缺点：需要的计算量大、推广性能低、误差估计特性差
  - 给定样本数量  $N$ 
    - $l$  必须足够大，以使其可以学习得到不同类之间的差异，以及相同类中模式的相似性
    - $l$  必须足够小，以使其不会将相同类的样本之间的差异也学出来。
    - 在实际中，一般取  $l < N/3$ 。
  - 选择“最好”的  $l$  个特征
    - 最好的特征具有：大的类间距离，小的类内方差。
- 特征选择的方法：假设检验法、可分性判别法、散布矩阵度量法，...





特征 $x_1$ 对分类有利，而特征 $x_2$ 则对分类不利

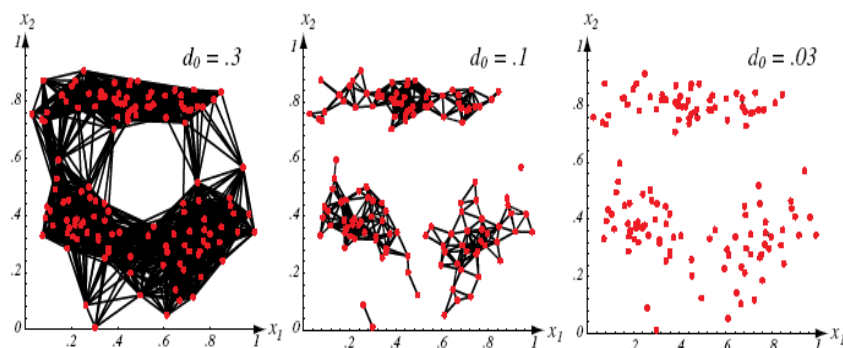
坏的、不坏的、以及好的  
特征选择



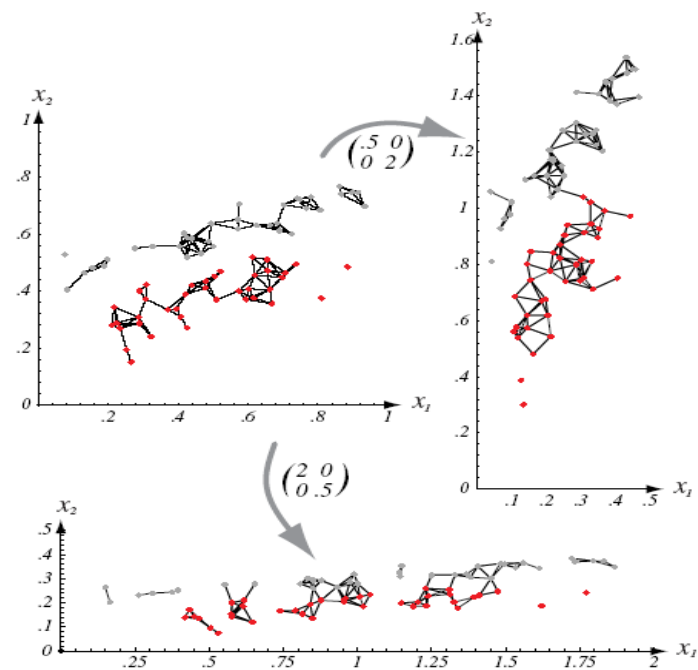
# 聚类：相似性度量

## ➤ 相似性度量

- 一般采用距离对相似性进行度量
- 要求所有选择的特征对相似性测度计算的贡献都是均衡的，没有那一个特征是绝对占优的



将距离小于阈值 $d_0$ 的样本连接后的结果：  
左 $d_0=0.3$ ，中 $d_0=0.1$ ，右 $d_0=0.03$



由于各个特征尺度的变化，引起了  
空间聚类结构的变化





# 聚类：相似性度量

- 明氏（Minkowski）距离

$$d(\mathbf{x}, \mathbf{v}) = \left[ \sum_{k=1}^d |x_k - v_k|^p \right]^{1/p}$$

- 马氏（Mahalanobis）距离

$$d(\mathbf{x}, \mathbf{v}) = \left[ (\mathbf{x} - \mathbf{v})^T S^{-1} (\mathbf{x} - \mathbf{v}) \right]^{1/2}$$

- 内积（角度距离）

$$s(\mathbf{x}, \mathbf{v}) = \frac{\mathbf{x}^T \mathbf{v}}{\|\mathbf{x}\| \|\mathbf{v}\|}$$

- Tanimoto 距离（0/1 二值特征向量情况）

$$d_T(\mathbf{x}, \mathbf{v}) = \frac{\mathbf{x}^T \mathbf{v}}{\|\mathbf{x}\| + \|\mathbf{v}\| - \mathbf{x}^T \mathbf{v}}$$



# 聚类：准则函数（目标函数）

## ➤ 聚类准则

- 聚类准则反映了我们对所期望的、数据集合内部隐含的聚类结构的判断与解释。
- 不同的聚类准则是对不同的数据相似性的一种反映。

## ➤ 聚类准则函数

- 是对数据集合聚类准则的先验认识的一种数学表达形式。
- 聚类准则函数也称为目标函数。

## ➤ 常见的一些聚类准则函数

- 平方误差和准则。
- 相关最小方差准则。
- 散布准则。

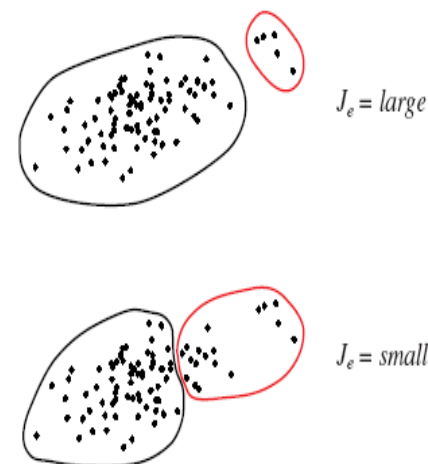


# 聚类准则函数： 1

平方误差和准则函数

$$J_e = \sum_{i=1}^c \sum_{k=1}^{n_i} \|x_{ik} - m_i\|^2$$

$$m_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik}$$



平方误差和准则的局限

相关最小方差准则函数

$$J_e = \frac{1}{2} \sum_{i=1}^c n_i S_i$$

$$S_i = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \|x_{ik} - x_{ij}\|^2$$

其中， $x_{ik}$ 是指第  $i$  类的第  $k$  个样本



# 聚类准则函数： 2

## 散布准则函数

**假设：**  $S_i$  是第  $i$  类的内散布矩阵， $S_w$  是类内散布矩阵， $S_B$  是类间散布矩阵， $S_T$  是总体散布矩阵

- 基于散布矩阵迹的准则（等价于误差平方和准则）

$$\text{tr}[S_w] = \sum_{i=1}^c \text{tr}[S_i] = \sum_{i=1}^c \sum_{k=1}^{n_i} \|x_{ik} - m_i\|^2 = J_e$$

- 基于散布矩阵行列式的准则

$$J_d = |S_w| = \left| \sum_{i=1}^c S_i \right|$$

- 基于散布矩阵不变量的准则

$$\text{tr}[S_w^{-1} S_B] = \sum_{i=1}^d \lambda_i \quad J_f = \text{tr}[S_T^{-1} S_B] = \sum_{i=1}^d \frac{1}{1 + \lambda_i} \quad \frac{|S_w|}{|S_T|} = \prod_{i=1}^d \frac{1}{1 + \lambda_i}$$



# 聚类：优化算法

1. 迭代优化算法。基于梯度下降的迭代优化算法。如k均值聚类

$$u_{ik} = \begin{cases} 1; & d_{ik} = \min_{1 \leq j \leq c} \{d_{jk}\} \\ 0; & otherwise \end{cases} \quad 1 \leq i \leq c \quad 1 \leq k \leq n$$

$$v_i = \frac{\sum_{k=1}^n u_{ik} x_k}{\sum_{k=1}^n u_{ik}} \quad 1 \leq i \leq c$$

2. 全局优化算法。

- 模拟退火算法
- 遗传算法
- 粒子群算法。



# 聚类：结构的有效性

- 聚类结构的有效性。在聚类算法获得了结果后，需要采用合适的检验方法检验其正确性。





# 聚类分析研究的一些热点

- 聚类算法的可伸缩性
- 处理噪声数据的能力
- 处理不同类型属性的能力
- 基于约束的聚类
- 高维聚类分析技术
- 对输入记录的顺序不敏感
- 发现任意形状的簇
- 用于决定输入参数的领域知识最小化
- 聚类算法对聚类复杂形状和类型的数据的有效性
- 可解释性和可用性

