上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

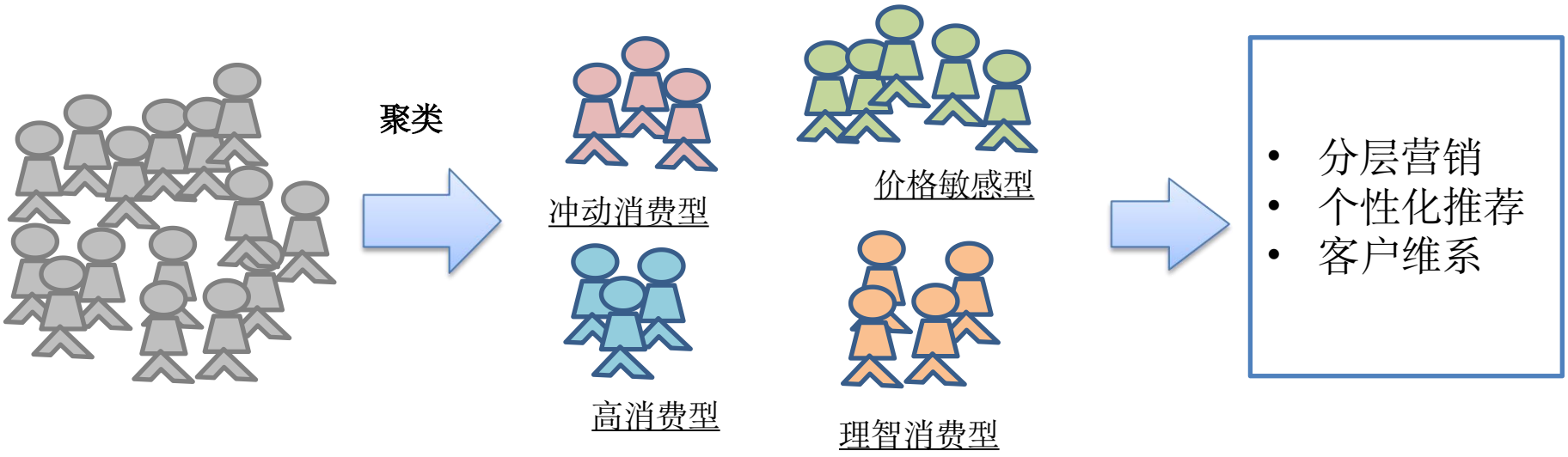
数据挖掘与商务分析

聚类分析

主讲教师：肖升生
xiao.shengsheng@shufe.edu.cn

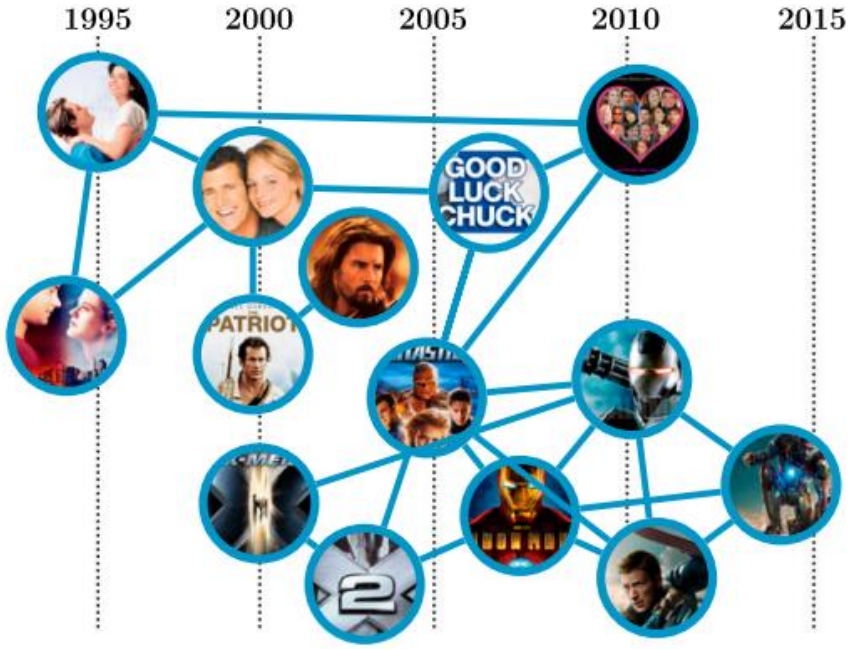
课程导入：客户细分

- 例子1：客户画像与客户细分



课程导入：相似电影的分析

- 例子2：电影分析

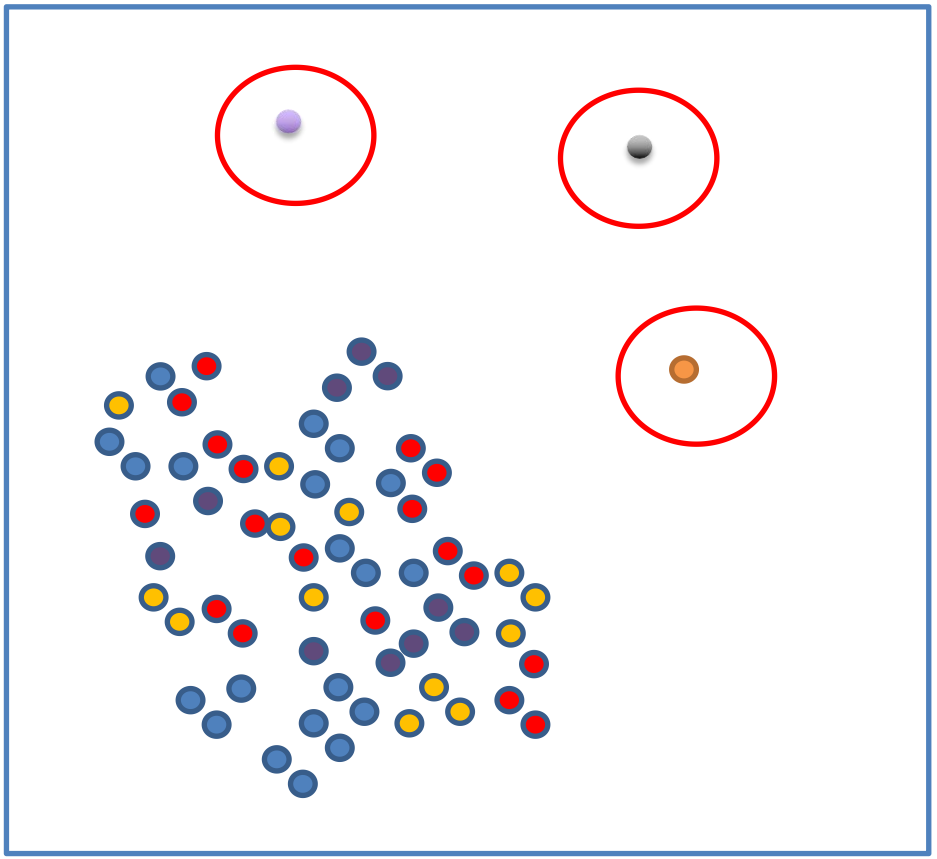


Notes. Movies are ordered from left to right by year of release. They are *Sleepless in Seattle* (1993), *While You Were Sleeping* (1995), *What Women Want* (2000), *The Patriot* (2000), *X Men* (2000), X2 (2003), *The Last Samurai* (2003), *Fantastic 4* (2005), *Good Luck Chuck* (2007), *Iron Man* (2008), *Valentine's Day* (2010), *Iron Man 2* (2010), *Captain America* (2011), and *Iron Man 3* (2013).

Yanhao Max Wei (2019) The Similarity Network of Motion Pictures. Management Science 66(4):1647-1671.

课程导入：异常检测

- 例子3：探测、发现孤立点、异常值
  - 保险欺诈
  - 信用卡欺诈
  - 故障检测
  - 系统健康监测



讲授提纲

01 聚类分析基本概念

02 聚类分析基本流程

03 聚类分析主要方法

04 聚类分析效果评估

05 商务案例分析

讲授提纲

01 聚类分析基本概念

02 聚类分析基本流程

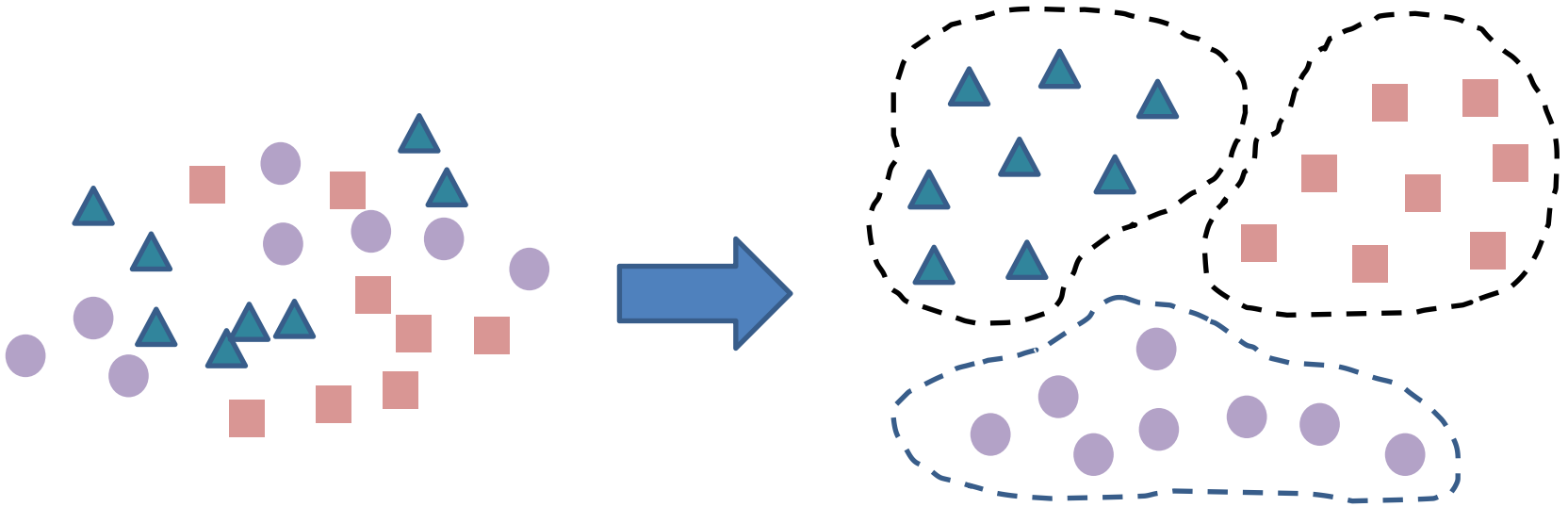
03 聚类分析主要方法

04 聚类分析效果评估

05 商务案例分析

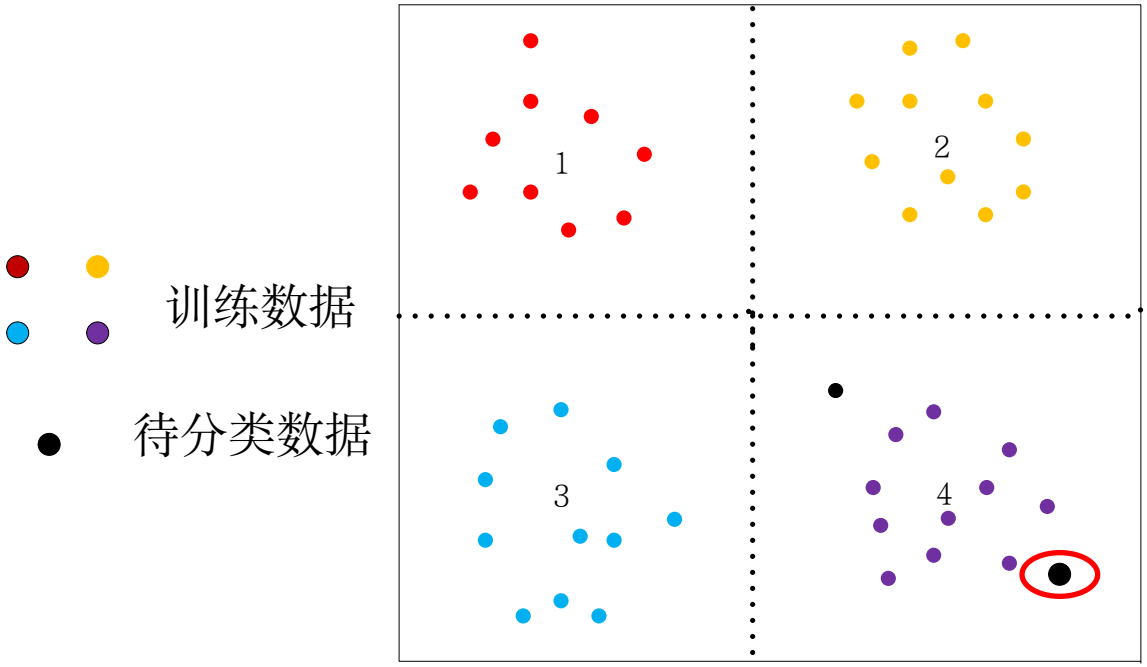
聚类分析的相关概念

- 聚类目标：将数据集中的样本划分为若干个通常不相交的子集（“簇”，Cluster）
- 聚类是一种无监督学习。



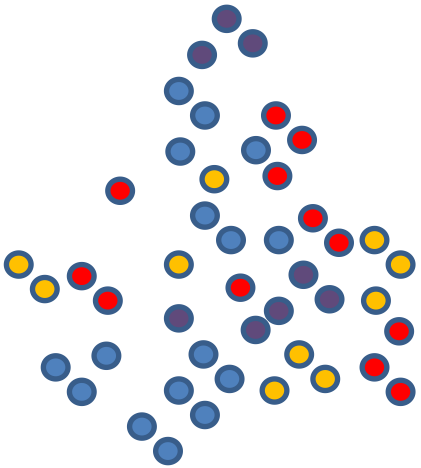
聚类分析的相关概念

- 有监督学习：有先验知识指导（标记样本，训练语料）



聚类分析的相关概念

- 无监督学习：无先验知识指导（无训练语料，无标记样本）
  - 把“对象”分成不同的类别，这些类不是事先给定的，而是直接根据数据的特征确定
  - 聚类中没有任何指导信息，完全按照数据的分布进行类别划分
  - 聚类的大小和结构都没有事先假定



？？

聚类分析的相关概念

- 聚类的基本原则：
  - 按数据的内在相似性将数据集划分为多个类别
  - 使类别内的数据相似度较大
  - 类别间的数据相似度较小

- 相似性度量方法
  - 基于距离的度量方法
  - 基于夹角余弦的度量方法
  - 基于相关系数的度量方法

聚类的原则--相似性度量

- 基于距离的度量方法

  - 欧氏距离（Euclidean distance）

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

q=2

  - 曼哈顿距离（Manhattan distance）

$$d(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$

q=1

  - 闵可夫斯基距离（Minkowski distance）

$$d(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^q\right)^{1/q}$$

聚类的原则--相似性度量

- 基于向量夹角的度量方法

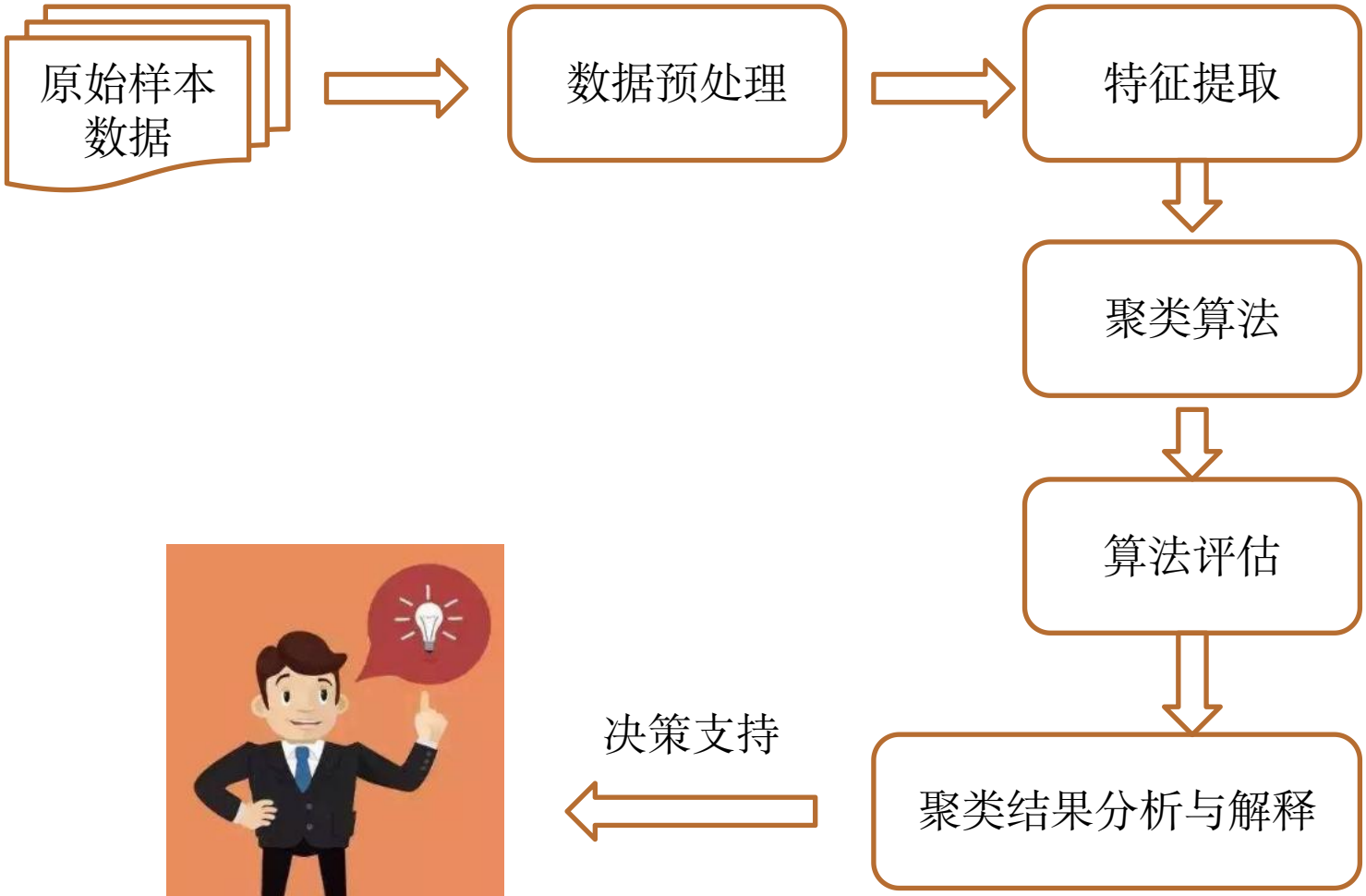
余弦相似度	$\cos\theta_{xy} = \dfrac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$

- 基于相关系数的度量方法

Pearson相关系数	$r_{xy} = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$

讲授提纲

聚类分析的基本流程



讲授提纲

聚类算法

基于划分的算法
(partitioning methods)

基于层次的算法
(hierarchical methods)

基于密度的算法
(density-based
methods)

基于划分的方法

- 划分聚类的核心思想：
  - 把相似的点划分为同一类，不相似的点划分到不同类
  - 划分聚类是聚类分析中最常用、最普遍的方法，简单易于使用，在实际中广泛运用

- 典型方法
  - K均值算法（k-means）
  - K中心点算法（k-medoid）

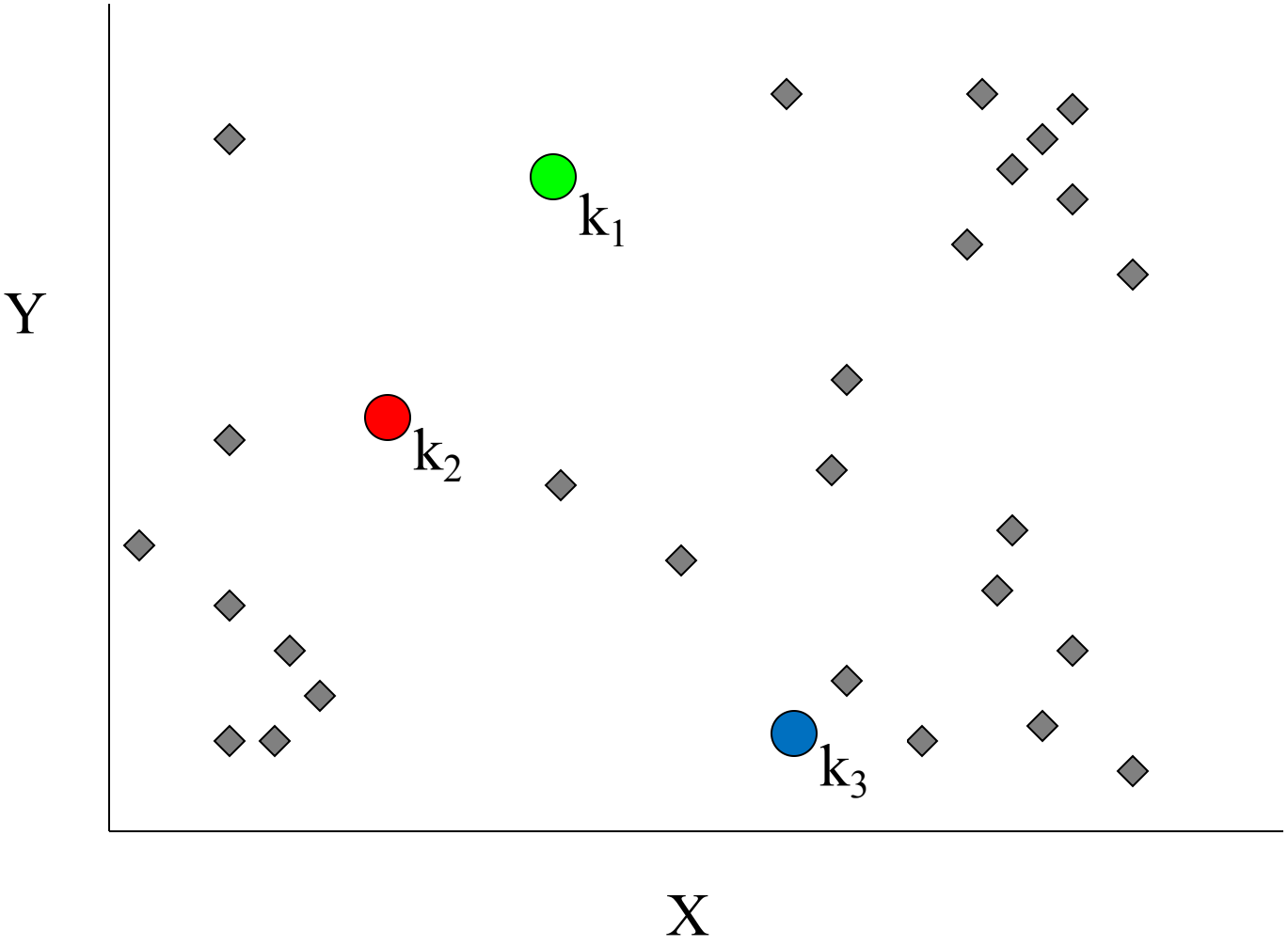
基于划分的方法： K-means

- 思路：根据样本点与子类中心的距离聚类，循环调整类别中心位置
- 具体步骤：

1. 设置初始子类数K，人为设置K个类别中心；
2. 根据样本和子类中心的距离进行类别划分，样本划分到距离最近的类别；
3. 重新计算当前每个子类的中心点（类别样本平均值）；
4. 在新的子类中心下继续进行类别划分；
5. 如果连续两次的类别划分结果不变则停止算法；否则循环2～5；

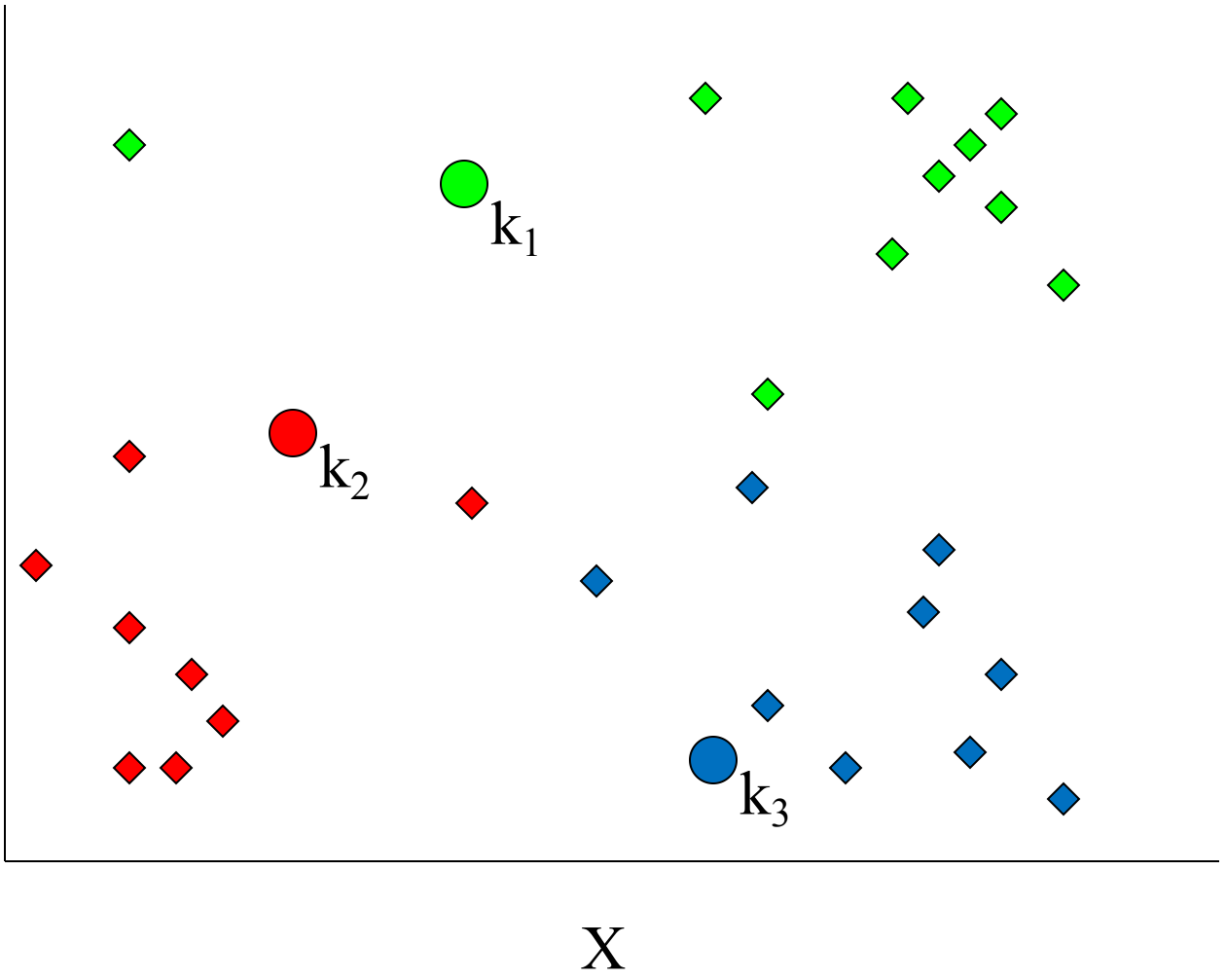
K-Means 聚类示例 (1)

随机选取三个初始中心点



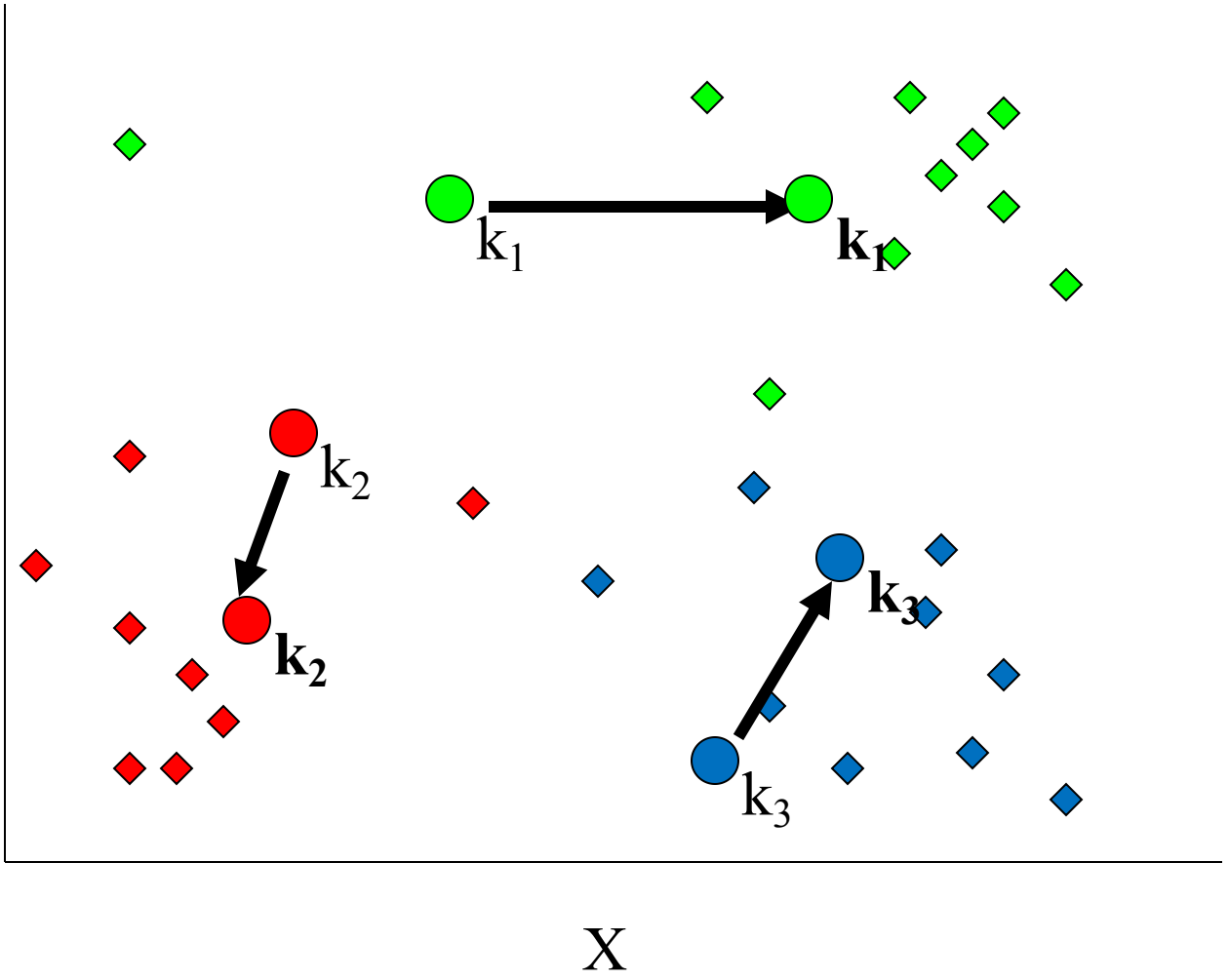
K-Means 聚类示例 (2)

将每个点分配到离它最近的簇中心



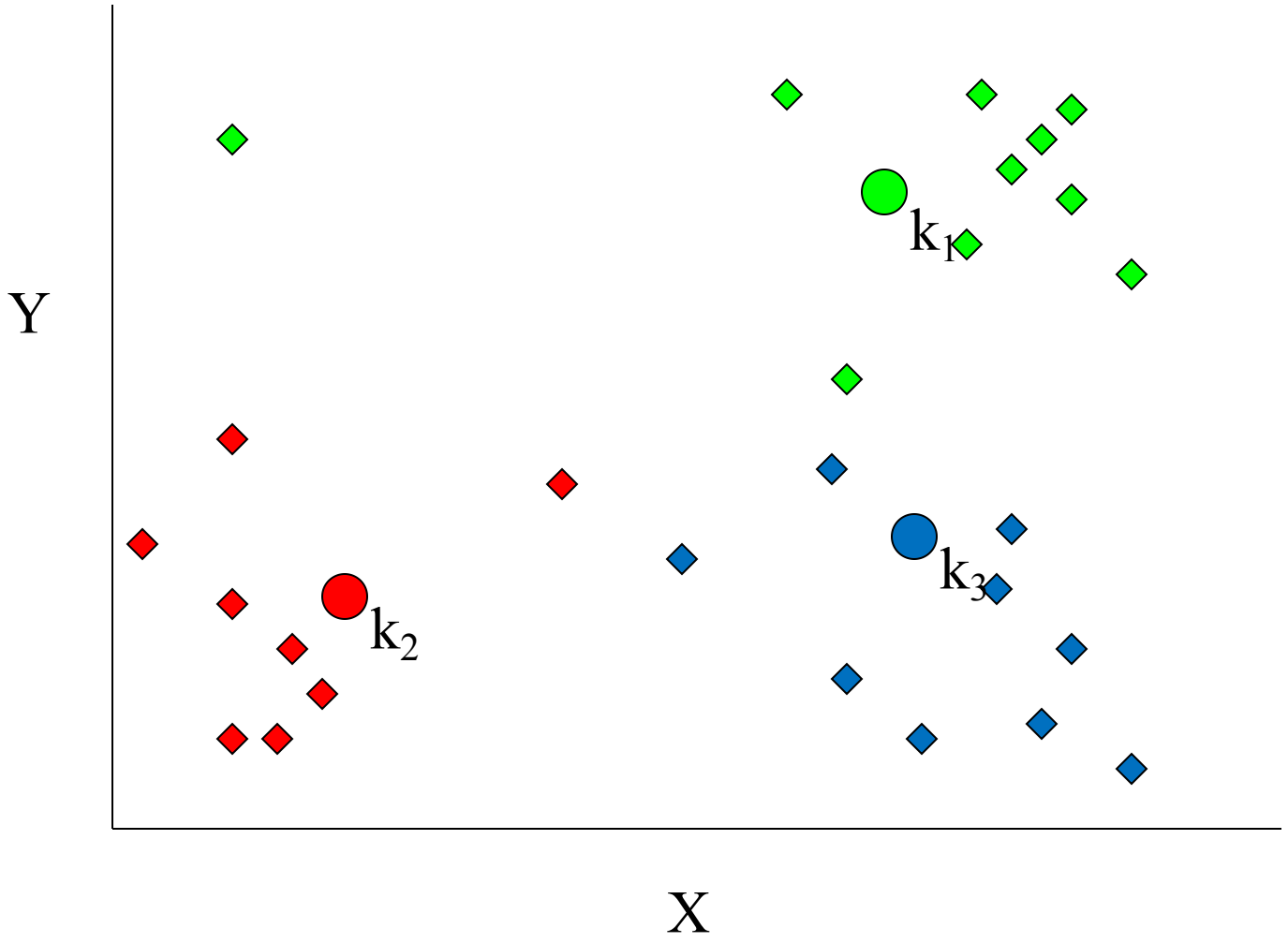
K-Means 聚类示例 (3)

移动簇中心位置到各个簇体的均值点



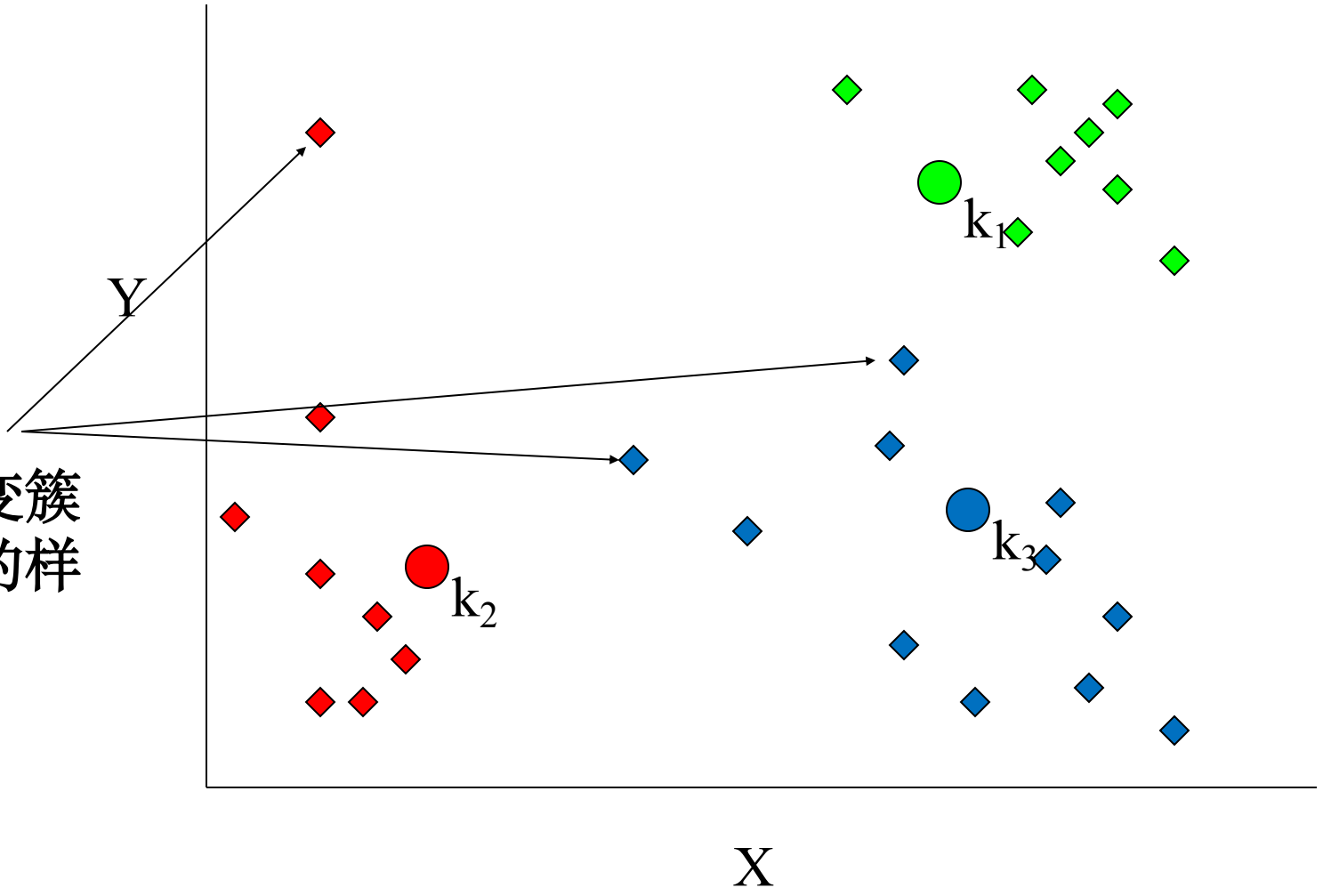
K-Means 聚类示例 (4)

重新分配数据点的簇体归属



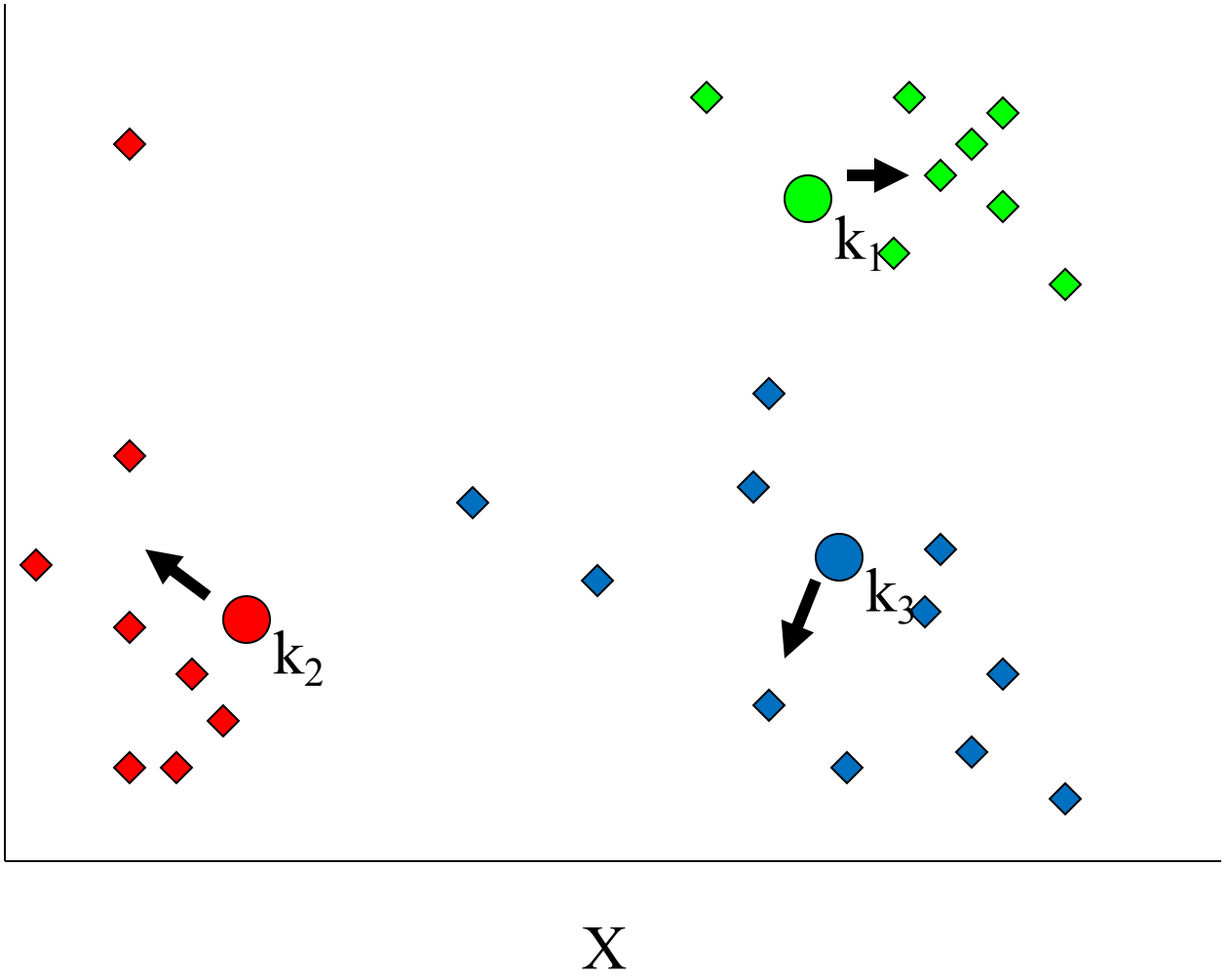
K-Means 聚类示例 (5)

三个改变簇体归属的样本点



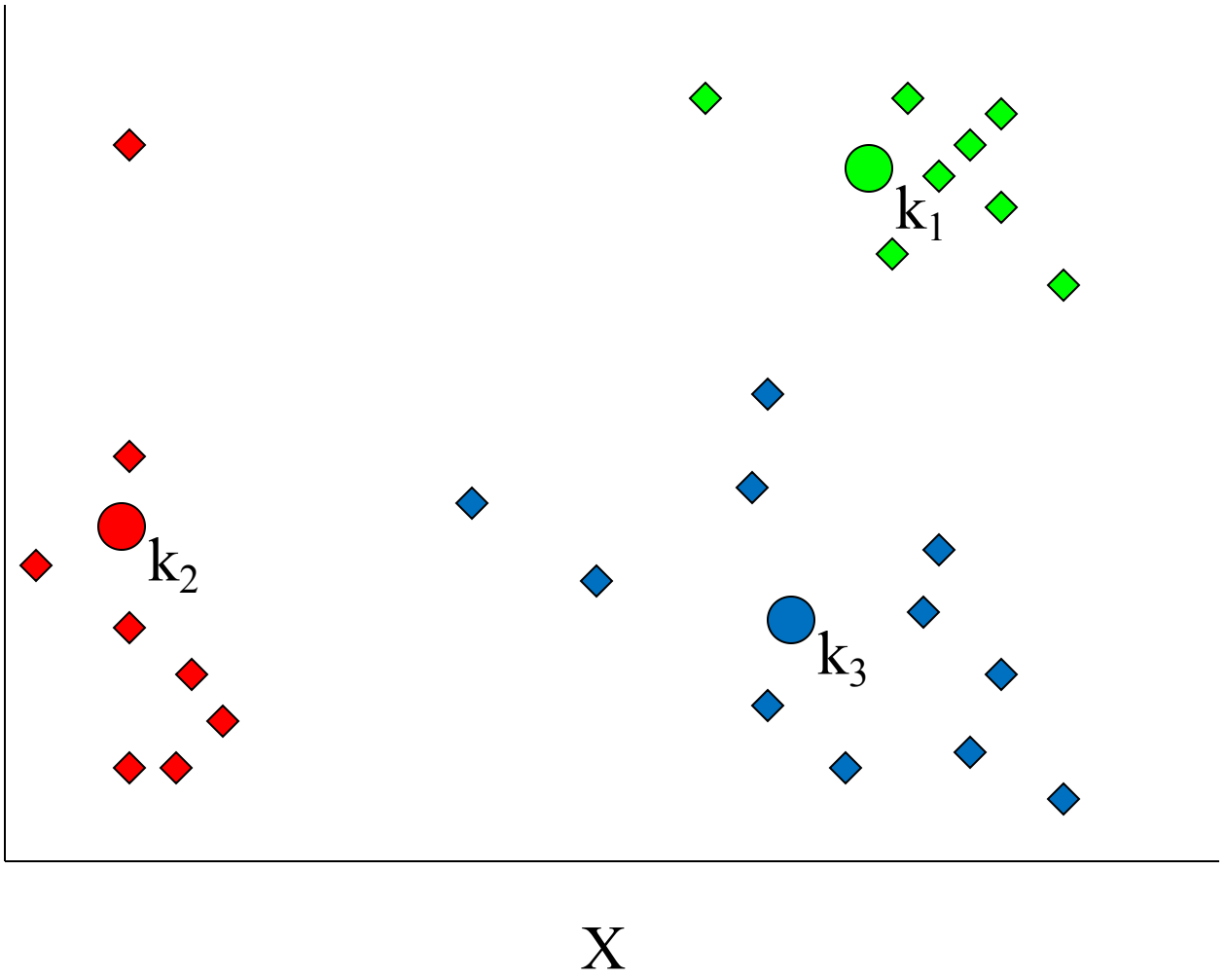
K-Means 聚类示例 (6)

重新计算
簇体均值点



K-Means 聚类示例 (7)

将簇中心移动到簇体均值点



K-means：如何确定最优K

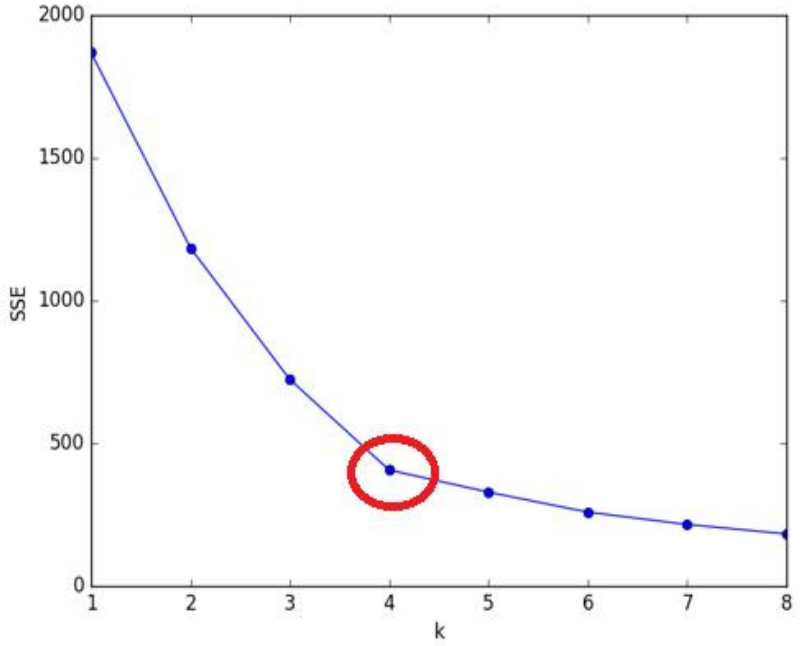
- **肘部法**
  - 核心指标：SSE(sum of the squared errors，误差平方和)

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

- 等式中各个变量的含义：
  - C_i是第i个簇
  - p是C_i中的样本点
  - m_i是C_i的质心（Ci中所有样本的均值）
  - SSE是所有样本的聚类误差，代表了聚类效果的好坏。

K-means：如何确定最优K

- 肘部法核心思想
  - 随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。
  - 当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减。



K-means：如何确定最优K

- 轮廓系数法
  - 使用轮廓系数来确定：选择使系数较大所对应的k值

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

- 等式中各个变量的含义：
  - a(i) :样本i的簇内不相似度。
  - b(i):样本i的簇间不相似度： bi =min{bi1, bi2, ..., bik}
  - s(i): 样本i的轮廓系数

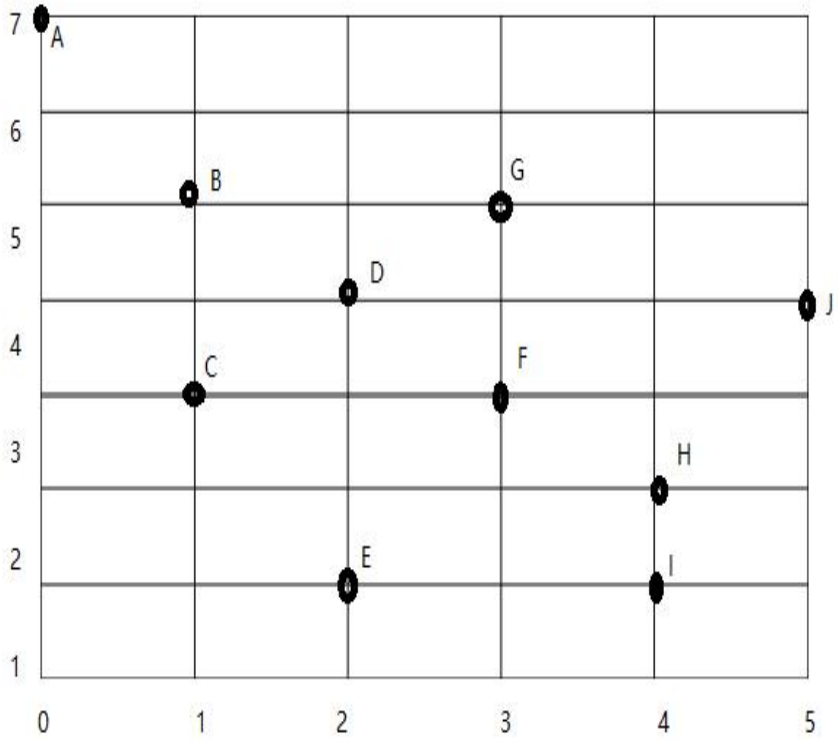
K-means 聚类方法总结

- **优点：**
  - 原理简单，实现容易
  - 容易解释
  - 计算时间短，速度快

- **缺点：**
  - 对初始值和异常值敏感
  - 结果不稳定 （受输入顺序影响）
  - 需要提前确定k值

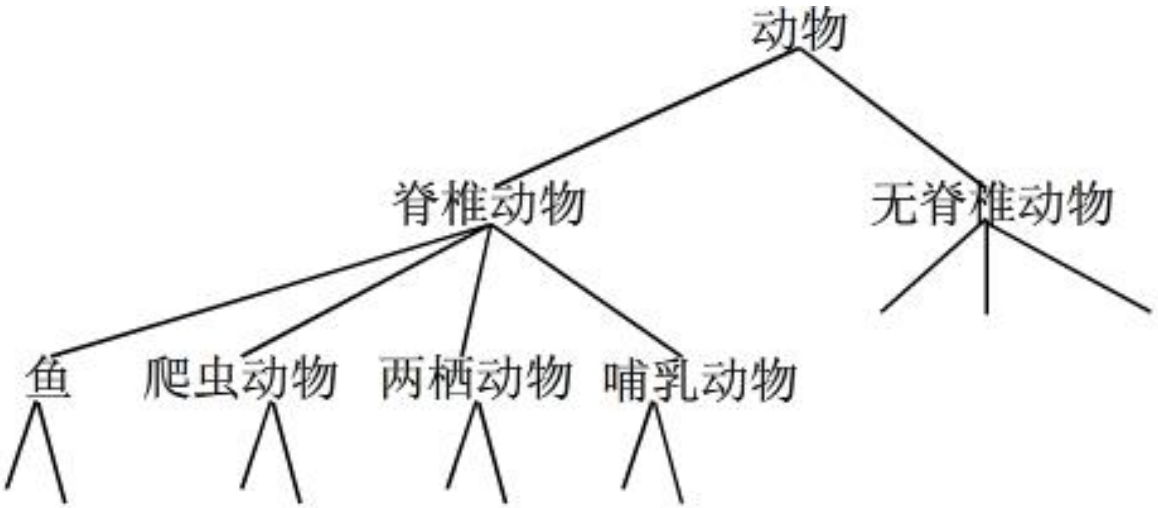
课堂练习：K-means

- 练习：请使用kmeans算法将下面数据分成三类。



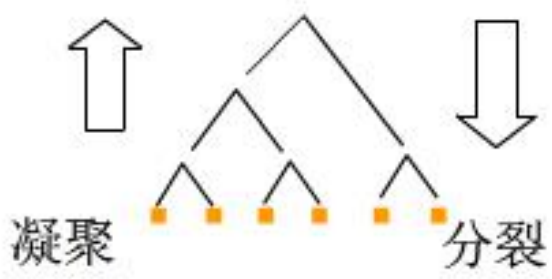
基于层次的聚类算法

- 层次聚类(Hierarchical Clustering)
  - 通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。
  - 在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。



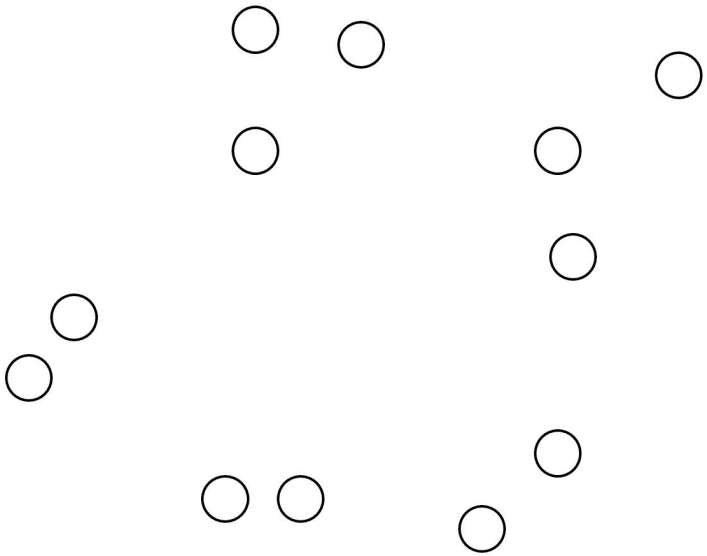
基于层次的聚类算法

- 层次聚类方法具体又可分为：
  - 凝聚的层次聚类：一种自底向上的策略，首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足。
  - 分裂的层次聚类：采用自顶向下的策略，它首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到达到了某个终结条件。



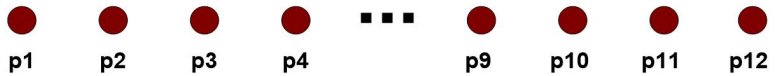
基于层次的聚类算法

- 一个示例：



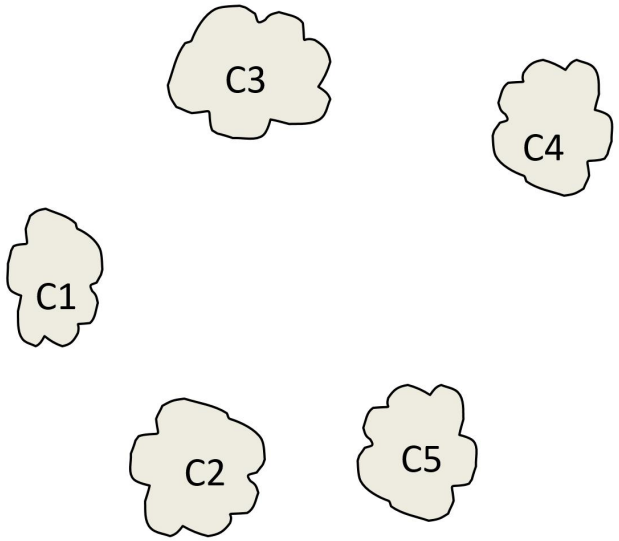
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



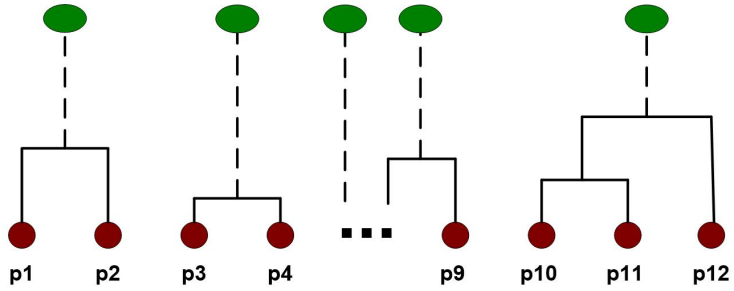
基于层次的聚类算法

- 一个示例：



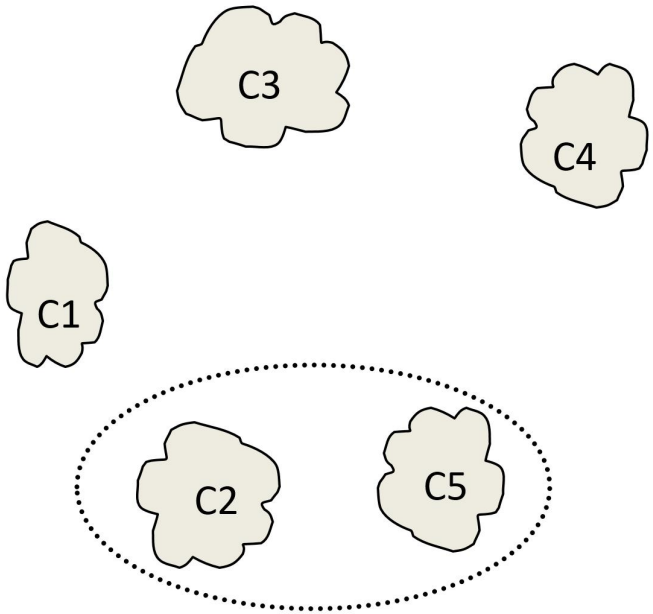
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



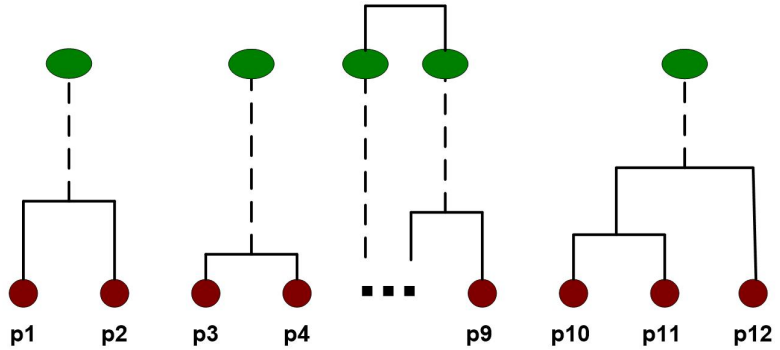
基于层次的聚类算法

- 一个示例：

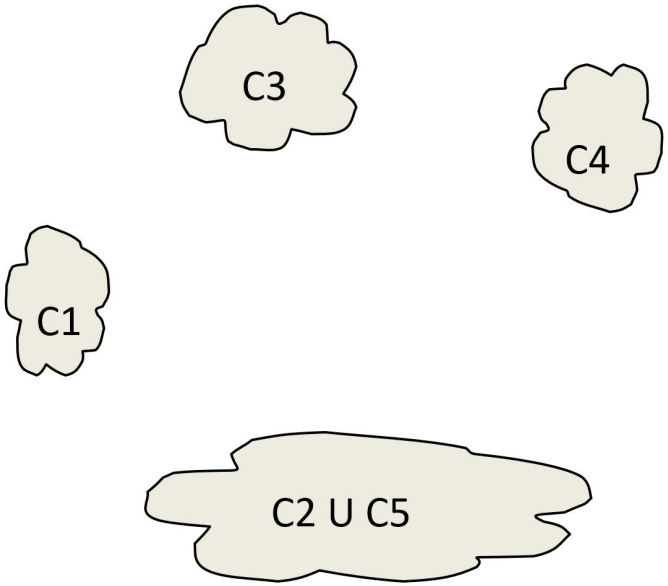


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

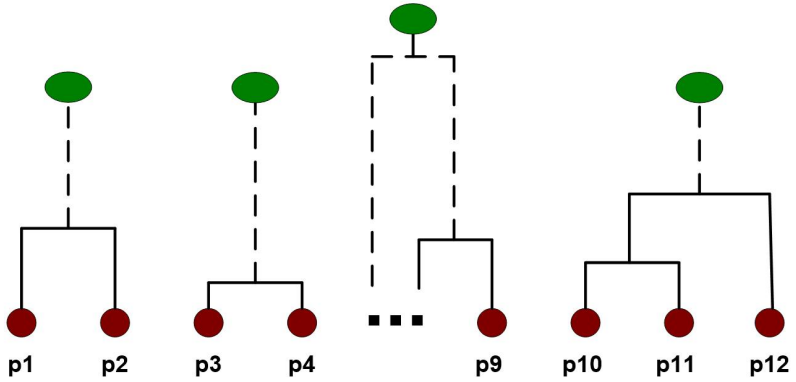


基于层次的聚类算法



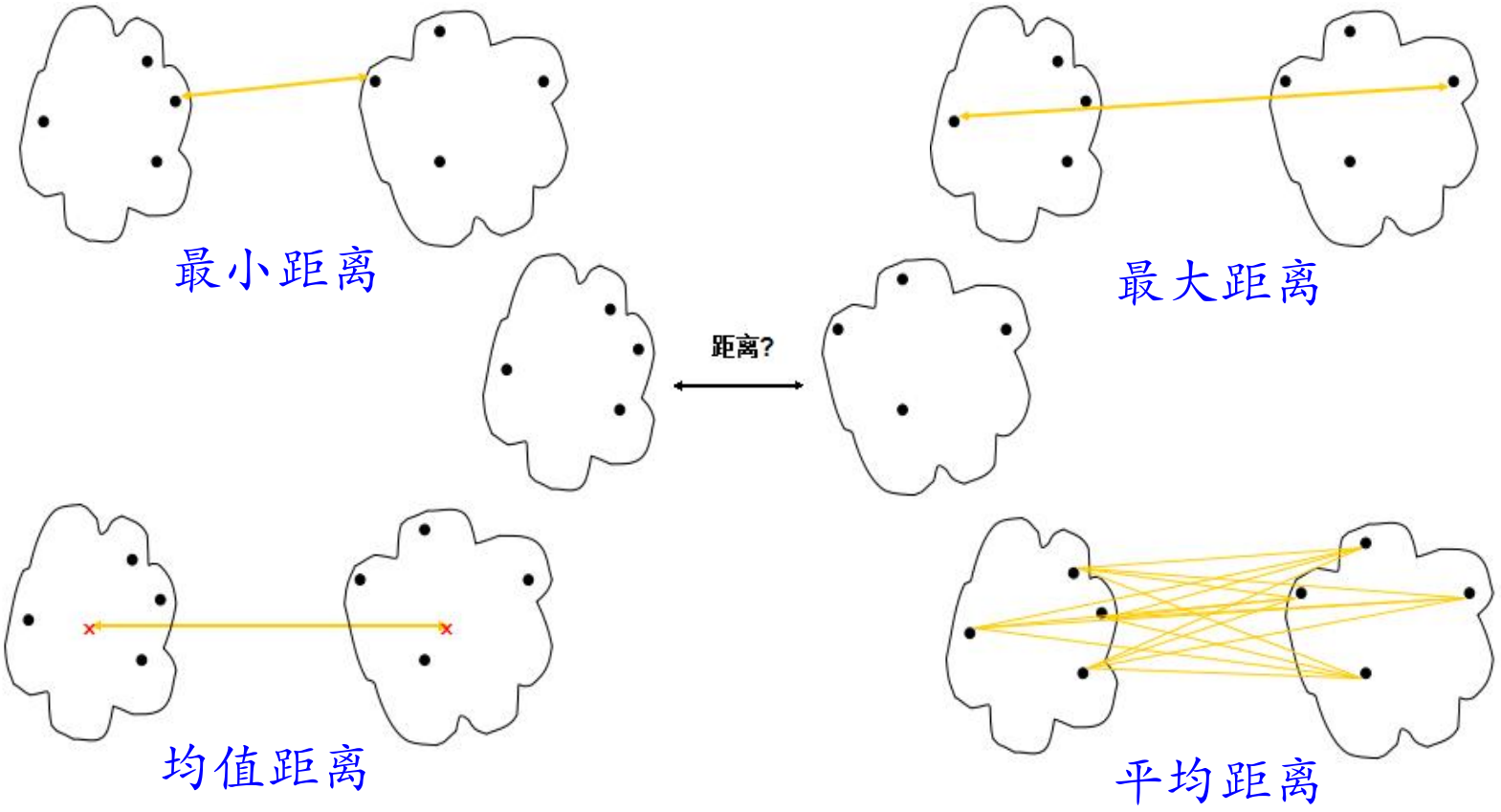
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



基于层次的聚类算法

- 簇间距离计算方法



基于密度的聚类方法

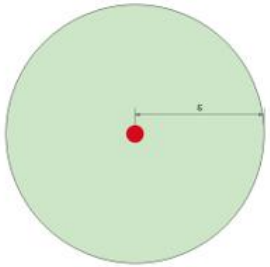
- **核心思想：**
  - 只要一个区域中的点的密度大于某个域值，就把它加到与之相近的聚类中去

- **特点：**
  - 可以过滤噪声和孤立点
  - 能克服基于距离的算法只能发现“类圆形”的聚类的缺点
  - 发现任意形状的类。

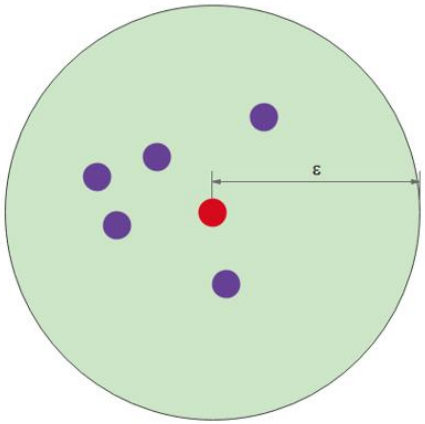
- **代表算法**
  - DBSCAN

基于密度的算法-DBSCAN

- **ε 邻域**
  - 给定对象半径为 ε 内的区域称为该对象的 ε 邻域



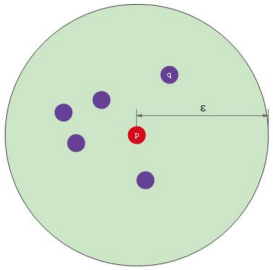
- **核心对象**
  - 在一个样本对象 C的 ε邻域中，
  - 有超过一定阈值 MinPts（最小数量）的样本对象分布，那么该样本对象 C就是核心对象



基于密度的算法-DBSCAN

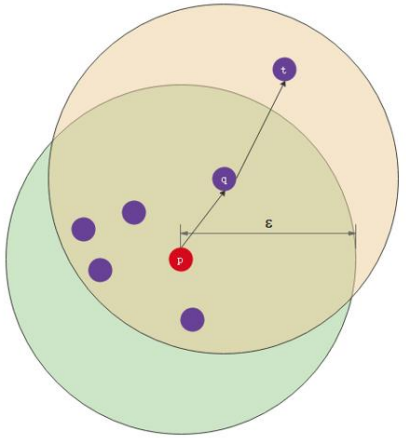
- **直接密度可达**：
  - 对于样本集合D，如果样本点q在p的 ε 邻域内，并且p为核心对象，那么对象q从对象p直接密度可达。

红色点 p 是核心对象，q 在其 ε-邻域 中，p 直接密度可达 q；



- **密度可达**：
  - 如果存在样本序列p1,p2,…,pt,满足q由pt直接密度可达，则称q由p1密度可达。
  - 此时序列中的传递样本p1,p2,…,pt均为核心对象，只有核心样本到其他对象才有密度直接可达。
  - 密度可达满足传递性

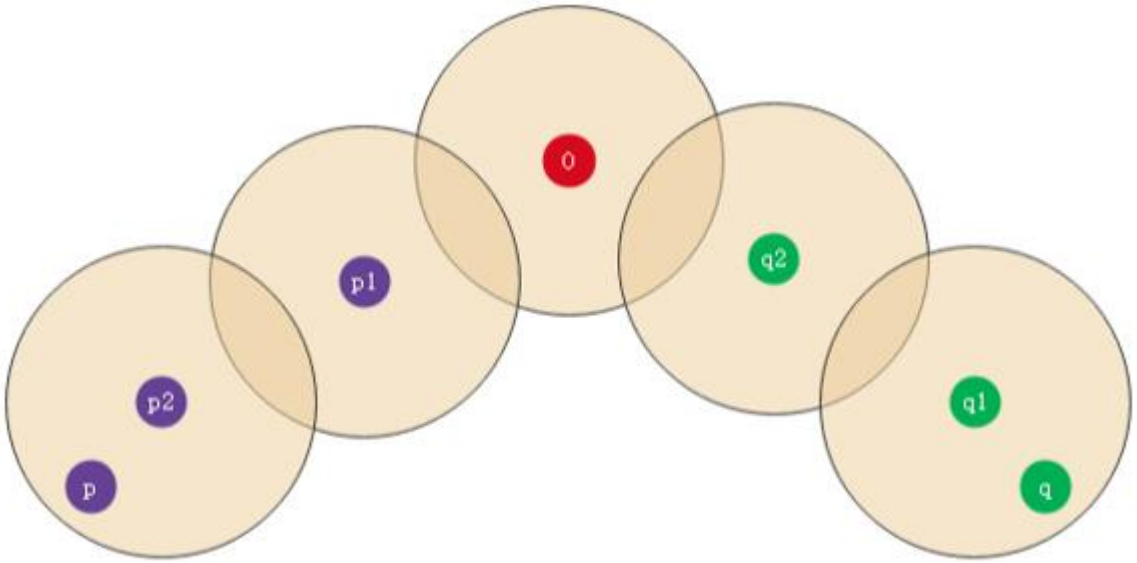
p 直接密度可达 q，q 直接密度可达 t，p 密度可达 t；



基于密度的算法-DBSCAN

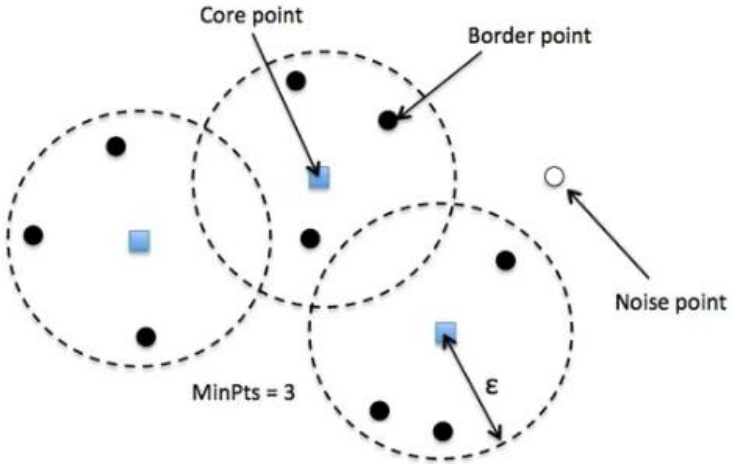
密度相连:

- 存在样本集合D中的一点o，如果对象o到对象p和对象q都是密度可达的，那么p和q密度相联。



基于密度的算法-DBSCAN

- 在DBSCAN算法中将数据点分为三类：
  - 核心点(core point)：在半径ε邻域含有超过MinPts数目的点；如果p(x)>=M，那么称x为X的核心点。
  - 边界点 (border point):在半径ε邻域内点的数量小于MinPts，但是落在核心点的邻域内的点。如果非核心点x的ε邻域中存在核心点，那么认为x为X的边界点。
  - 噪音点(noise point)：集合中除了边界点和核心点之外的点都是噪音点。



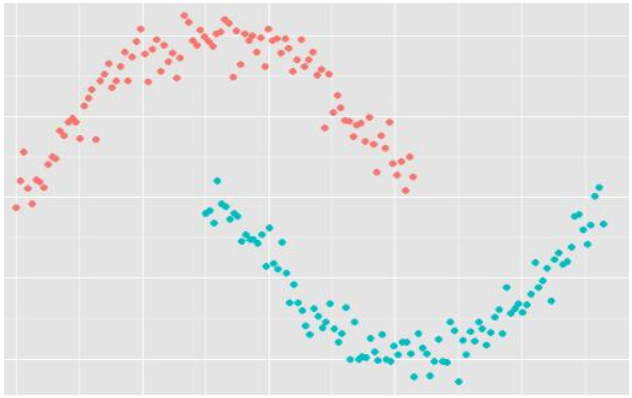
基于密度的算法-DBSCAN

- **算法步骤（在已知ε和MinPts的前提下）**
  - 任意选择一个未被标记的点，判断它是否为核心点。如果是，在该点周围建立一个类，否则，设定为噪音点。
  - 遍历其他点，直到建立一个簇；把直接密度可达的点加入到簇中，接着把密度可达的点也加进来。如果标记为噪音的点被加进来，修改状态为边界点。
  - 重复步骤1和2，直到所有的点满足在类中（核心点或边界点）或者为噪音点。

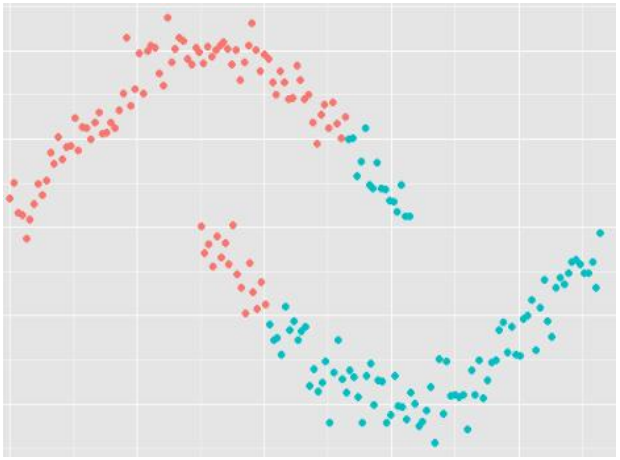
基于密度的算法-DBSCAN

优点：

- 与K-means方法相比，DBSCAN不需要事先知道要形成的簇类的数量。
- 与K-means方法相比，DBSCAN可以发现任意形状的簇类。
- DBSCAN能够识别出噪声点。对离群点有较好的鲁棒性，甚至可以检测离群点。
- DBSCAN对于数据库中样本的顺序不敏感，即Pattern的输入顺序对结果的影响不大。
- 对于处于簇类之间边界样本，可能会根据哪个簇类优先被探测到而其归属有所摆动。



DBSCAN聚类效果



k-means的聚类效果

基于密度的算法-DBSCAN

缺点：

- DBSCAN不能很好反映高维数据。
- 由于DBSCAN算法使用了全局性表征密度的参数，若不同簇类的样本集密度相差很大，则DBSCAN的聚类效果很差。
- 调参相对于传统的K-Means之类的聚类算法稍复杂，主要需要对距离阈值 ε ，邻域样本数阈值MinPts联合调参，不同的参数组合对最后的聚类效果有较大影响。
- 如果数据集是稀疏的，并且数据集是凸数据集，那么用DBSCAN的聚类效果不一定好。

讲授提纲

聚类效果评估

- 聚类效果评估，亦称为聚类“有效性指标”：
  - “簇内相似度”（intra-cluster similarity）高
  - “簇间相似度”（inter-cluster similarity）低

- 聚类效果评估方法：
  - 外部指标 (external index)评估法
  - 内部指标 (internal index)评估法

聚类效果评估

- 外部指标评估法
  - 有监督的方法，需要基准数据。
  - 用一定的度量评判聚类结果与基准数据的符合程度。
  - 基准是一种理想的聚类，通常由专家构建。

聚类效果评估

		算法聚类	
		同簇	不同簇
外部标准	同簇	a	c
	不同簇	b	d

- 其中a+b+c+d=M； M=N*(N-1)/2，N表示X中所有样本的总数
- 常见的评估指标包括：
  - Rand指数 （Rand Index, RI）

    RI= （a+d） /M
  - Jaccad 系数 （Jaccard Coefficient, JC）

    J=a/(a+b+c)
  - FM指数 （Fowlkes and Mallows Index, FMI）

    $$FMI = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

[0,1]区间内，越大越好.

聚类结果评估－内部指标

- 内部指标评估法：
  - 无监督的方法，无需基准数据
  - 根据类内聚集程度和类间离散程度判定
- 根据聚类结果的簇划分$C=\{C_1, C_2, \cdots C_k\}$，
  - 定义簇C内样本间的平均距离

$$\mathrm{avg}(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \leqslant i < j \leqslant |C|} \mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  - 定义样本间的最远距离

$$\mathrm{diam}(C) = \max_{1 \leqslant i < j \leqslant |C|} \mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  - 定义样本C_i，C_j最近样本间的距离

$$d_{\min}(C_i, C_j) = \min_{\boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j} \mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  - 定义簇C_i与C_j中心点间的距离

$$d_{\mathrm{cen}}(C_i, C_j) = \mathrm{dist}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$$

聚类结果评估－内部指标

- DB指数（Davies-Bouldin Index, DBI）

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

越小越好.

- Dunn指数（Dunn Index, DI）

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

越大越好.

聚类结果评估－内部指标

- 轮廓系数（silhouette coefficient）：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

- a(i)的值反映i所属的簇的紧凑性。该值越小，簇越紧凑
- b(i)的值捕获i与其他簇的分离程度。b(i)的值越大，i与其他簇越分离
- S(i)接近1时，包含i的簇是紧凑的，并且i远离其他簇，说明样本i聚类合理
- S(i)接近-1时这意味在期望情况下，i距离其他簇的对象比距离与自己同在簇的对象更近，说明样本i更应该分类到另外的簇
- 若si 近似为0，则说明样本i在两个簇的边界上

小结

- 聚类分析的基本概念
- 聚类分析的常见方法
  - K-means
  - 层次聚类方法
  - 基于密度的聚类方法
- 聚类结果的评估

讲授提纲