聚类分析

本章主要内容

1. 聚类分析基本原理及层次聚类法
2. 聚类分析中的各种距离
3. 层次聚类法：SPSS实现
4. K均值聚类法及SPSS实现

本章主要内容

1. 聚类分析基本原理及层次聚类法
2. 聚类分析中的各种距离
3. 层次聚类法：SPSS实现
4. K均值聚类法及SPSS实现

聚类分析基本原理

- 聚类分析，这里叫聚类，不是分类，有人可能会感到奇怪，明明是对样本进行分类，为啥不能叫分类呢?
- 因为在数据挖掘中，分类一般指带有标签的样本数据的分类，而聚类解决的是没有标签的样本数据的分类。
- 例如，一个公司有不同的部门，研发、产品、营销等，这些不同的部门会将员工分成不同的类别，这个叫作分类，因为每个人身上有部门这个标签，根据标签可以将不同的人分到不同的类别。
- 但是，这个公司的员工可以根据自己的兴趣爱好、家乡等，自发形成不用的圈子，这个就叫作聚类，这里的兴趣爱好、家乡等可以理解为人与人之间的“距离”，根据这个“距离”，不同的人会融入不同的群体。
- 聚类也是如此，没有标签，根据样本之间的距离将样本分成不同的类别。

聚类分析基本原理

- 例如，之前讲过的一个案例，用逻辑回归预测企业员工是否离职，样本数据如下。

satisfaction_level	last_evaluation	number_project	……	Work_accident	left（类别标签）
0.38	0.53	2	……	0	1
0.8	0.86	5	……	0	0
0.11	0.88	7	……	0	0
0.72	0.87	5	……	0	1
0.37	0.52	2	……	0	1
0.41	0.5	2	……	0	0

- 样本数据：特征X、类别标签y，需要用分类算法解决。
- 常见的分类算法：K近邻算法、逻辑回归、决策树、SVM等。

聚类分析基本原理

- 例如，用户分群管理，已知的样本数据只有用户的特征，如首次消费时间、最近一次消费时间、消费总金额等，需要根据用户的特征将用户分为不同的类别（类别事先未知）

用户编码	首次消费时间	最近一次消费时间	消费总金额	消费总次数
1	2016/6/25	2016/6/26	20000	3
2	2016/6/26	2016/6/26	50000	3
3	2016/6/13	2016/6/19	108000	4
4	2016/6/16	2016/6/16	30000	3
5	2016/6/27	2016/6/27	100000	3
6	2016/6/22	2016/6/22	15000	3

聚类分析基本原理

- 再比如，下表是同一批客户对经常光顾的五座商厦在购物环境和服务质量两方面的平均评分。现希望根据这批数据将五座商厦分类。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

这类问题需要用聚类分析来解决，常见的聚类方法有层次聚类法（系统聚类法）、K均值聚类法。

层次聚类法

层次聚类法，hierarchical clustering method，也叫系统聚类法，是使用最多的一种聚类方法。

计算步骤：

1、开始每个样本自成一类，有几个样本就有几个类。

2、计算各个类之间的距离，每次合并距离最近的两个类。

3、重复步骤2，直到所有样本都合并为一个类。

距离的计算

假设两个样本点$(x_1, y_1), (x_2, y_2)$，它们的距离公式为：

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

案例：商厦分类

- 下表是同一批客户对经常光顾的五座商厦在购物环境和服务质量两方面的平均评分。现希望根据这批数据将五座商厦分类。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

层次聚类法计算过程

1、开始每个样本自成一类，有几个样本就有几个类。

5个样本，各自成为一类，得到5个类：A，B，C，D，E

2、计算各个类之间的距离，每次合并距离最近的两个类。

由于最开始是每个样本自成一类，所以计算类间距离相当于计算每个样本之间的距离。

例如，计算类A和B之间的距离

$$d(A,B)=\sqrt{(73-66)^2+(68-64)^2}=8.06$$

按照类似的方式，可以将所有类之间的距离都计算出来。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

$$D_0=\begin{bmatrix} & A & B & C & D & E \\ A & 0 & & & & \\ B & 8.06 & 0 & & & \\ C & 17.80 & 25.46 & 0 & & \\ D & 26.91 & 34.66 & 9.22 & 0 & \\ E & 30.41 & 38.21 & 12.81 & 3.61 & 0 \end{bmatrix}$$

层次聚类法计算过程

从距离矩阵看出，D,E这两个类之间的距离最小，

因此将D，E合并为一个新类，记为CL4，CL4={D,E}

合并之后，剩下4个类：A，B，C，CL4

3、重复步骤2，直到所有样本都合并为一个类。

接着，计算这4个类之间的距离。

说明：这里按照最短距离法计算类间距离。

从距离矩阵看出，A,B之间的距离最小。

$$D_0 = \begin{bmatrix} & A & B & C & D & E \\ A & 0 & & & & \\ B & 8.06 & 0 & & & \\ C & 17.80 & 25.46 & 0 & & \\ D & 26.91 & 34.66 & 9.22 & 0 & \\ E & 30.41 & 38.21 & 12.81 & 3.61 & 0 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} & A & B & C & CL4 = \{D, E\} \\ A & 0 & & & \\ B & 8.06 & 0 & & \\ C & 17.80 & 25.46 & 0 & \\ CL4 = \{D, E\} & 26.91 & 34.66 & 9.22 & 0 \end{bmatrix}$$

因此将A，B合并为一个新类，记为CL3，CL3={A,B}

合并之后，剩下3个类：CL3，C，CL4

层次聚类法计算过程

接着计算各类之间的距离矩阵

$$D_2 = \begin{bmatrix} & CL3 & C & CL4=\{D,E\} \\ CL3 & 0 & & \\ C & 17.80 & 0 & \\ CL4=\{D,E\} & 26.91 & 9.22 & 0 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} & A & B & C & CL4=\{D,E\} \\ A & 0 & & & \\ B & 8.06 & 0 & & \\ C & 17.80 & 25.46 & 0 & \\ CL4=\{D,E\} & 26.91 & 34.66 & 9.22 & 0 \end{bmatrix}$$

从距离矩阵看出，C,CL4之间的距离最小。

因为将C，CL4合并为一个新类，记为CL2，CL2={C,CL4}

合并之后，剩下两个类：CL3，CL2，接着计算各类之间的距离矩阵

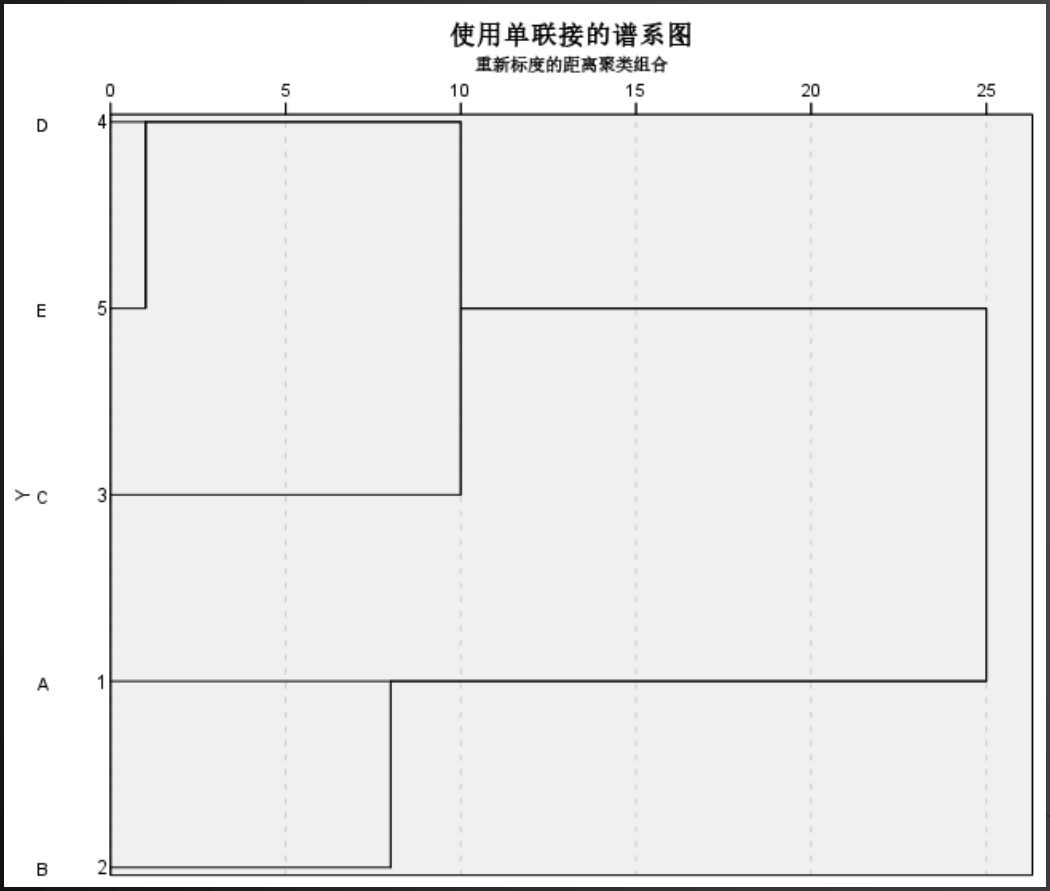
最后，将CL3和CL2合并为CL1。

至此，并类完成。

$$D_2 = \begin{bmatrix} & CL3 & CL2 \\ CL3 & 0 & \\ CL2 & 17.80 & 0 \end{bmatrix}$$

谱系聚类图

将上述聚类过程用一个图来表示，即谱系聚类图，如下图所示。



编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

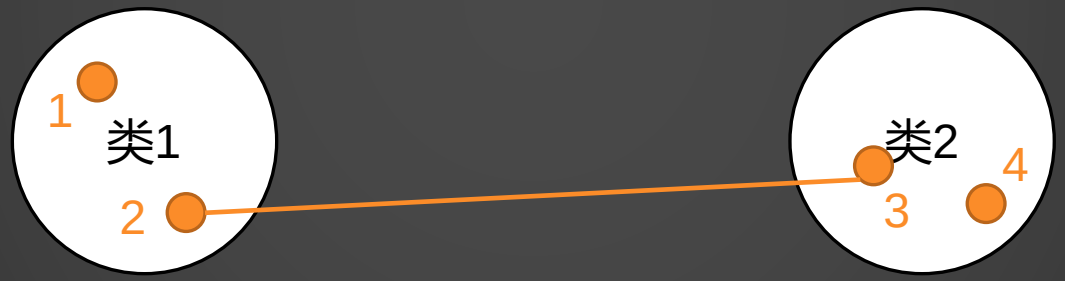
本章主要内容

1. 聚类分析基本原理及层次聚类法
2. 聚类分析中的各种距离
3. 层次聚类法：SPSS实现
4. K均值聚类法及SPSS实现

聚类分析中的各种距离

从前面的聚类过程可以看到，聚类分析中的距离有两种：

- 样本间的距离：欧氏距离、平方欧式距离、Block距离、Chebychev距离、闵可夫斯基距离等。
- 类间的距离：最短距离法、最长距离法、Ward距离、重心法、类平均法等。



样本间的距离

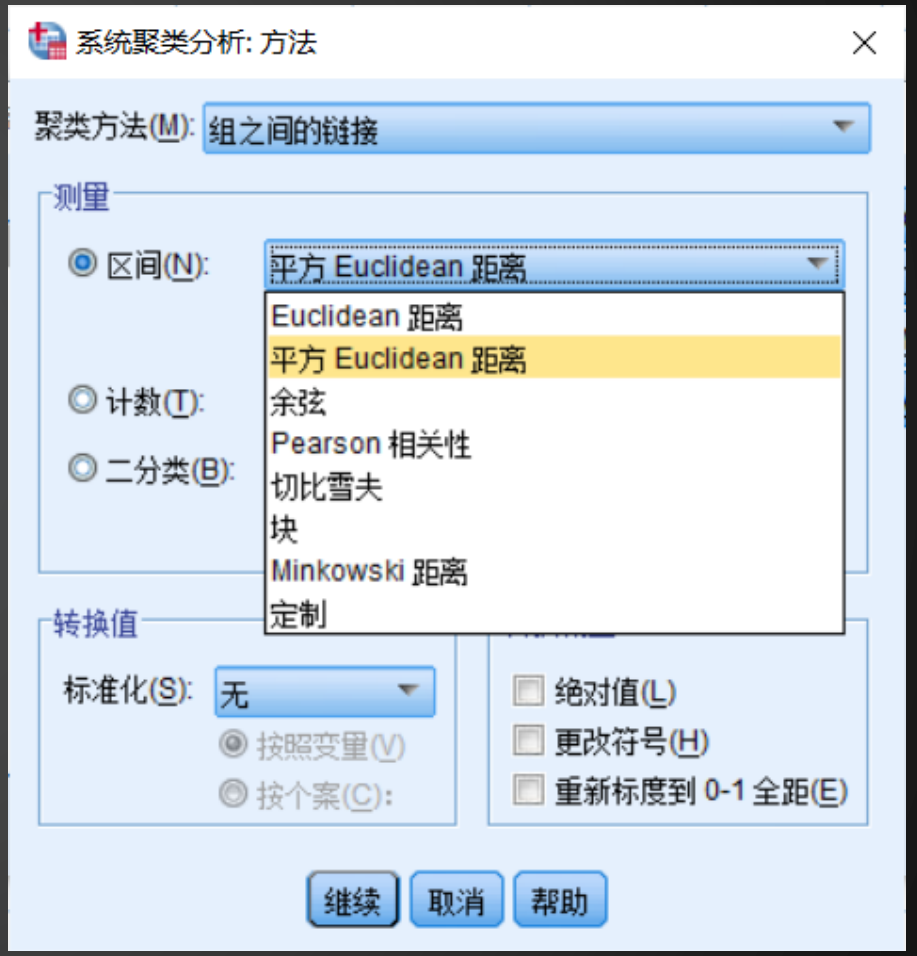
假设有两个点$A=(x_1,y_1),B=(x_2,y_2)$

- 欧氏距离：$d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
- 平方欧式距离：$d=(x_1-x_2)^2+(y_1-y_2)^2$
- Block距离：$d=|x_1-x_2|+|y_1-y_2|$
- Chebychev距离：$d=\max\{|x_1-x_2|,|y_1-y_2|\}$
- Minkowski距离：$d=\sqrt[q]{|x_1-x_2|^q+|y_1-y_2|^q}$

当$q=2$时，就是欧式距离

当$q=1$时，就是Block距离

当$q\to\infty$时，就是Chebychev距离



类间距离

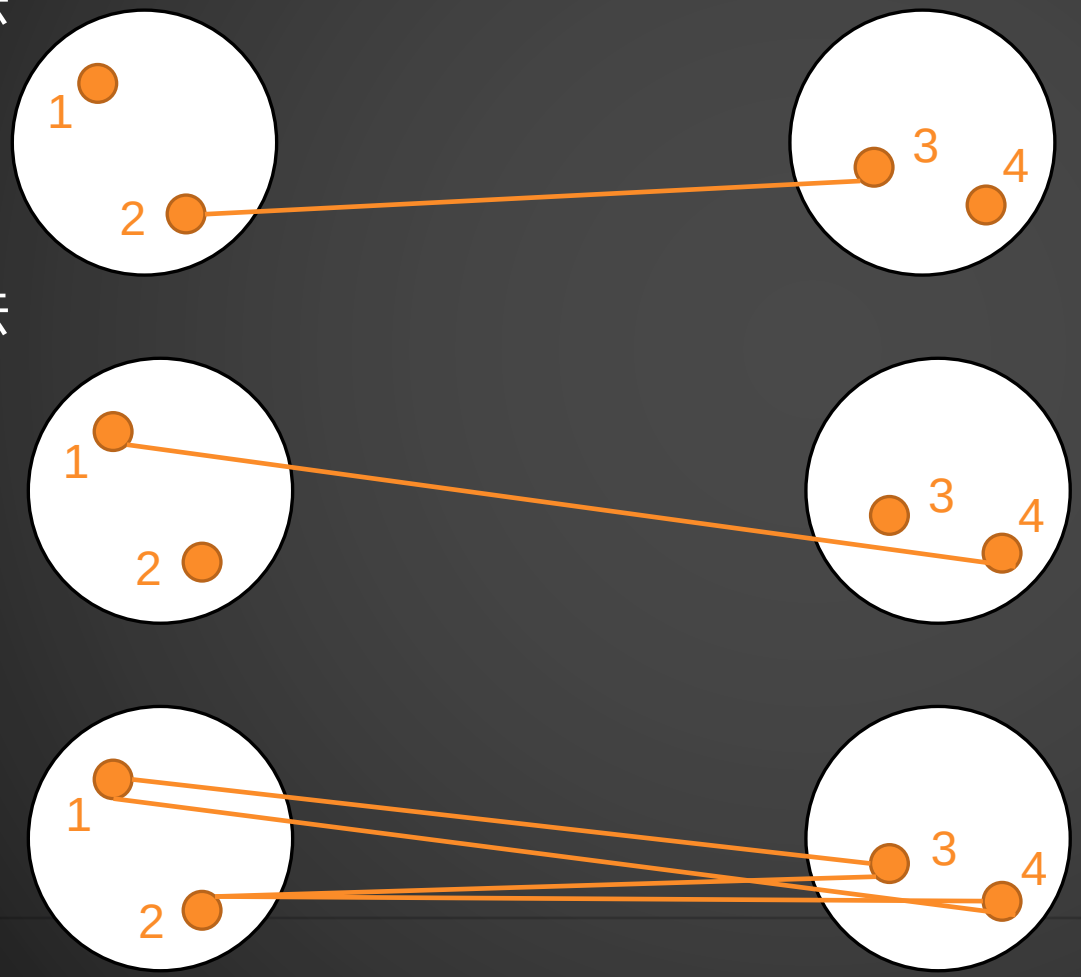
计算类间距离(与上面介绍的样本间距离不同)的方法有很多，不同方法会得到不同的聚类结果 。

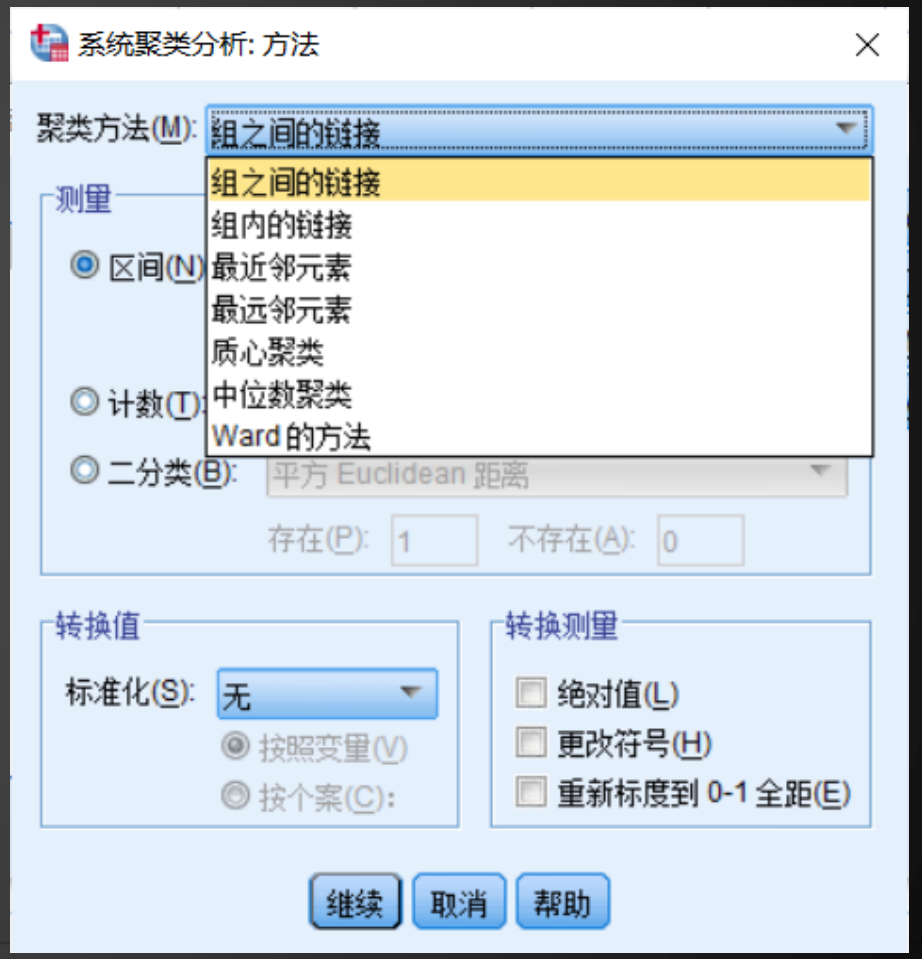
常用的计算类间距离的方法有：

- 最短距离法（single linkage method）
- 最长距离法（complete linkage method）
- Ward距离：离差平方法（Ward）
- 重心法（centroid method）
- 类平均法（group average method）

类间距离

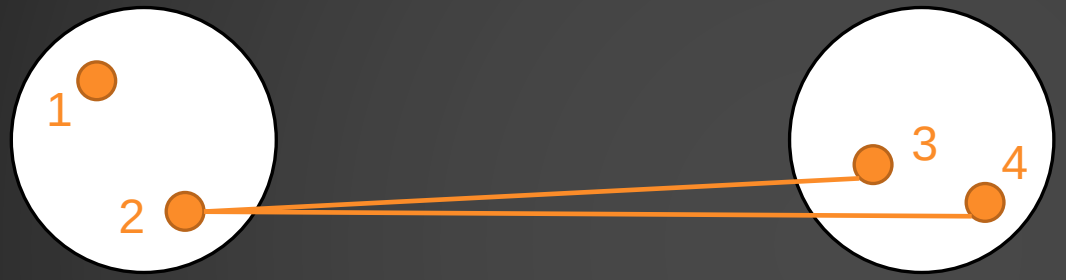
- 最短距离法
- 最长距离法
- 类平均法



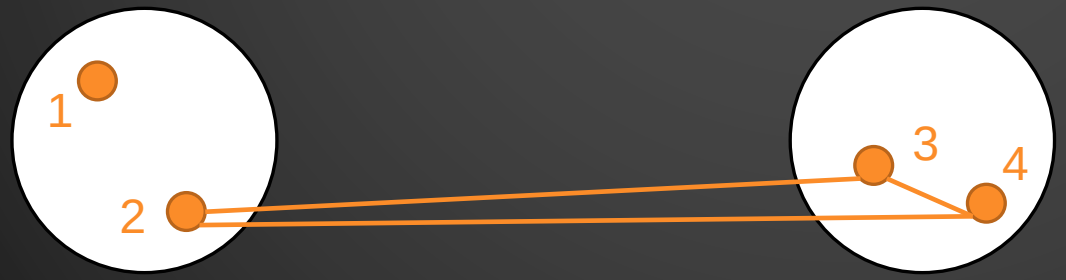


类间距离

- 组之间的链接：考虑样本与类中每个个体距离的平均值，防止极端值的影响。



- 组内的链接：考虑样本与类中每个个体距离，及类中样本距离的平均值，考虑到内部差异性。



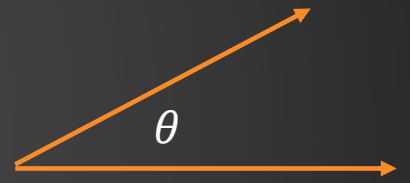
指标间的距离：夹角余弦

聚类分析不仅可以对样本进行分类，还可以对指标进行分类，测度指标相似的方法有两种：夹角余弦和相关系数。

例如，两个指标$X_i = (x_{1i}, x_{2i}, \cdots, x_{ni})$和$X_j = (x_{1j}, x_{2j}, \cdots, x_{nj})$

- 夹角余弦

$$C_{ij} = \frac{X_i \cdot X_j}{|X_i| \cdot |X_j|} = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\sqrt{(\sum_{k=1}^{n} x_{ki}^2)(\sum_{k=1}^{n} x_{kj}^2)}}$$



$$cos\theta = \frac{a \cdot b}{|a| \cdot |b|}$$

- 相关系数：相关系数就是数据标准化后的夹角余弦。

$$R_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{X}_i)(x_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{X}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{X}_j)^2}}$$

本章主要内容

1. 聚类分析基本原理及层次聚类法
2. 聚类分析中的各种距离
3. 层次聚类法：SPSS实现
4. K均值聚类法及SPSS实现

层次聚类法：SPSS实现

主要步骤：

1. 读取数据
2. SPSS菜单：【分析】-【分类】-【系统聚类】
3. 选择参与聚类的变量、聚类方法（类间的距离）、距离计算方式
4. 解释分析结果

本章主要内容

1. 聚类分析基本原理及层次聚类法
2. 聚类分析中的各种距离
3. 层次聚类法：SPSS实现
4. K均值聚类法及SPSS实现

K均值聚类法

K均值聚类，英文名称是k-means clustering，所以叫作K均值聚类法。

这里的k是指聚类类别数，例如，如果想将样本分为3类，则k=3，如果分为6类，则k=6。

计算步骤

1、确定聚类类别数、初始聚类中心

2、通过计算每一个样本与聚类中心之间的距离，将样本归到与之距离最近的类中

3、重新计算每个类的聚类中心，重复这样的过程，直到每个样本都被归入类中

K均值聚类法计算过程

1、确定聚类类别数、初始聚类中心

- 确定聚类类别数：假设分为2类，即k=2
- 选取初始聚类中心：B=(66,64)、D=(91,88)

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

K均值聚类法计算过程

2、通过计算每一个样本与聚类中心之间的距离，将样本归到距离最近的类中

样本A：离B最近，将其归入聚类中心B所在的类

样本C：离D最近，将其归入聚类中心D所在的类

$$D_0 = \begin{bmatrix} & B & D \\ A & 8.06 & 26.91 \\ C & 25.46 & 9.22 \end{bmatrix}$$

3、重新计算每个类的聚类中心，重复这样的过程，直到每个样本都被归入类中

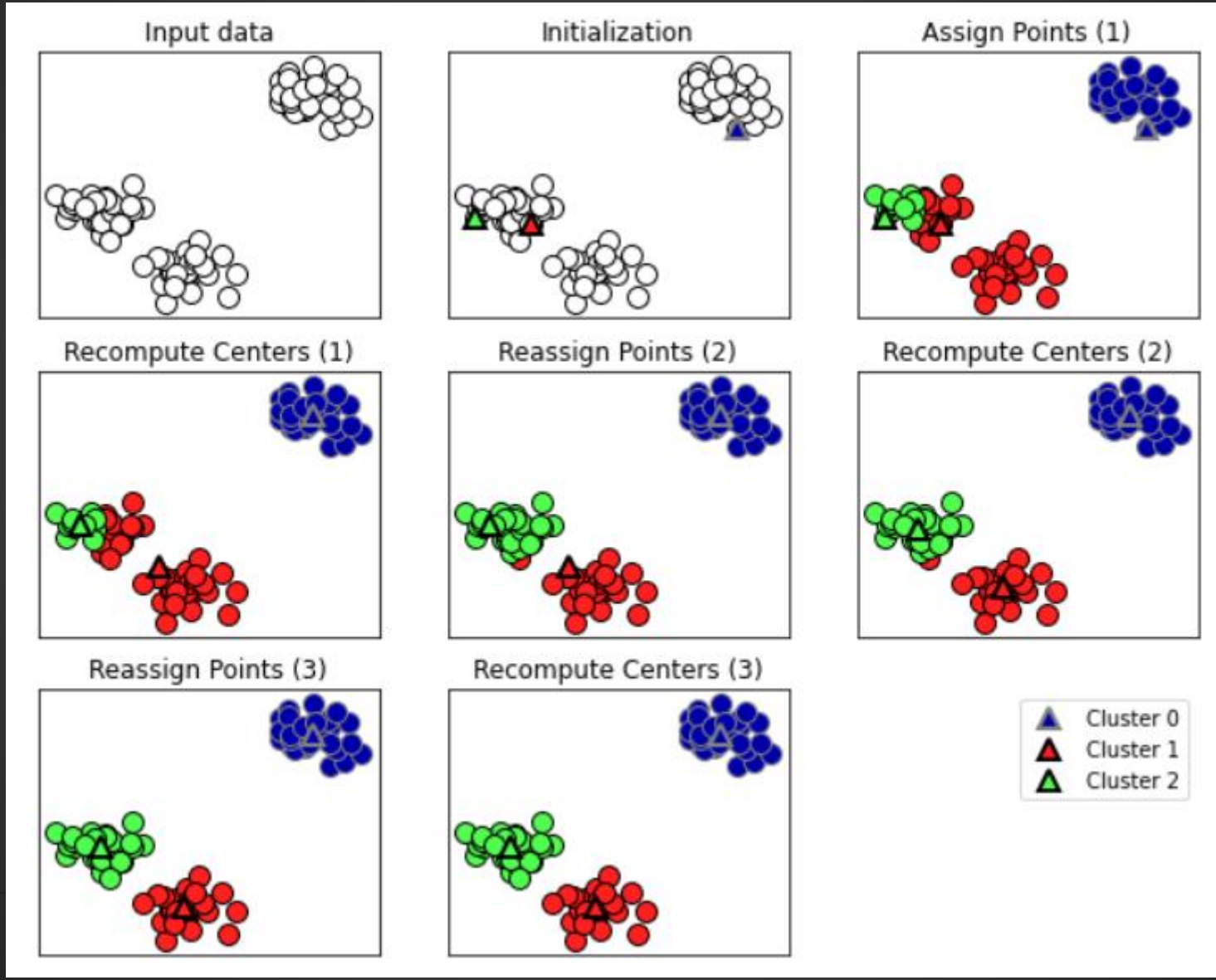
第一个聚类中心：B =(66,64)，变成{A,B}= $(69.5,66)$

第二个聚类中心：D =(91,88)，变成{C,D} = $(87.5,85)$

样本E：离聚类中心{C,D}最近，将其归入聚类中心{C,D}所在的类，聚类中心变成{C,D,E}

至此，得到两类，C1={A，B}，C2={C,D,E}，若要分成三类也是如此。

K均值聚类法流程示意图



K均值聚类法：SPSS实现

主要步骤

1. 数据读取
2. SPSS菜单：【分析】-【分类】-【K-均值聚类】
3. 选择参与聚类的变量、聚类数（K值）、初始聚类中心等
4. 解释分析结果

案例：商厦分类

- 下表是同一批客户对经常光顾的五座商厦在购物环境和服务质量两方面的平均评分。现希望根据这批数据将五座商厦分类。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90