多元统计

陈崇双

西南交通大学数学学院统计系

ccsmars@swjtu.edu.cn

2018-2019学年

截止至2018.06


快速检索
标准检索
专业检索
作者发文检索
科研基金检索
句子检索
来源期刊检索
1.输入检索控制条件：
期刊年期：从 不限 年 到 不限 年
指定期：
请输入
更新时间：
不限
来源期刊：
输入期刊名称，ISSN,CN均可
模糊
来源类别：
EI来源期刊
支持基金：
输入基金名称
模糊
作者
输入作者姓名
精确
作者单位：
输入作者单位，全称、简称、曾用名均可
模糊
2.输入内容检索条件：
关键词
聚类分析
词频
并且包含
输入检索词
词频
精确
或者
篇名
聚类分析
词频
并且包含
词频
精确
仅限优先出版论文
中英文扩展检索
在结果中检索
检索文献
定制本次检索式
3.您可以按如下文献分组排序方式选择文献：(分组只对前4万条记录分组,排序只在800万条记录以内有效)
文献分组浏览：学科类别 期刊名称 研究资助基金 研究层次 文献作者 作者单位 中文关键词 发表年度 不分组
"紫色"刊名为"中国知网"独家出版刊物
文献排序浏览：发表时间 相关度 被引频次 下载频次
列表显示
每页记录数：10 20 50
找到 1,034 条结果 浏览1/52
1 2 3 4 5 6 7 8 9 后页
全选
清除
存盘
定制

计算机软件及计算机应用(237)	自动化技术(118)
电力工业(96)	轻工业手工业(57)
公路与水路运输(44)	地质学(44)
环境科学与资源利用(43)	数学(38)
化学(37)	电信技术(35)
建筑科学与工程(30)	宏观经济管理与可持续发展(26)
矿业工程(25)	互联网技术(25)
生物学(24)	石油天然气工业(21)
机械工业(21)	园艺(19)
中药学(18)	地球物理学(18)
工业通用技术及设备(18)	农业基础科学(17)
自然地理学和测绘学(16)	农作物(15)
航空航天科学与工程(14)	企业经济(14)
水利水电工程(13)	安全科学与灾害防治(12)

1. 聚类分析
   - 基本思想
   - 相似性度量
   - 类和类的特征
   - 系统聚类法
   - K-均值聚类
   - 分类数的确定
   - 有序样本聚类
   - 延伸阅读
   - 作业

第一节：基本思想

主要内容：聚类分析的背景，定义

背景

- “物以类聚、人以群分”，“道不同不相为谋”

背景

- “物以类聚、人以群分”，“道不同不相为谋”
- 在一些社会经济问题中，我们面临的往往是比较复杂的研究对象，如果能把相似的样品（或指标）归成类，处理起来就大为方便。

背景

- “物以类聚、人以群分”，“道不同不相为谋”
- 在一些社会经济问题中，我们面临的往往是比较复杂的研究对象，如果能把相似的样品（或指标）归成类，处理起来就大为方便。
- 粗糙集理论认为，知识就是对于对象进行分类的能力。

背景

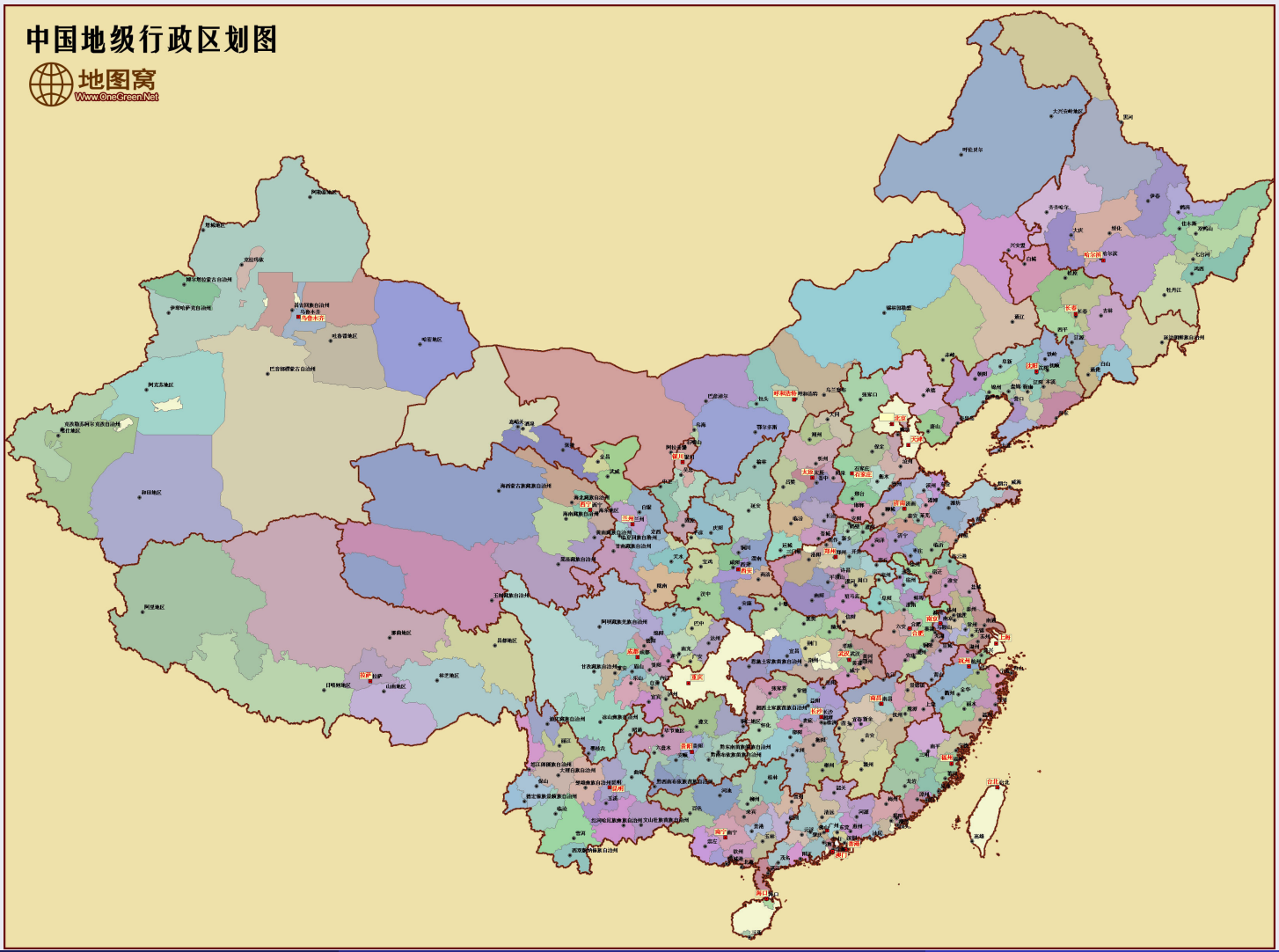
- “物以类聚、人以群分”，“道不同不相为谋”
- 在一些社会经济问题中，我们面临的往往是比较复杂的研究对象，如果能把相似的样品（或指标）归成类，处理起来就大为方便。
- 粗糙集理论认为，知识就是对于对象进行分类的能力。
- 分类的依据或者准则？

背景

- “物以类聚、人以群分”，“道不同不相为谋”
- 在一些社会经济问题中，我们面临的往往是比较复杂的研究对象，如果能把相似的样品（或指标）归成类，处理起来就大为方便。
- 粗糙集理论认为，知识就是对于对象进行分类的能力。
- 分类的依据或者准则？
- 靠经验等做定性处理，有主观性和任意性。特别是对于多指标分类，定性分析很难准确分类。

例1

对中国的市（约*660*个）进行分类：可按自然条件（降雨量、温度、海拔、地形等）；也可按经济水平（人均*GDP*、收入等）。



例2

将11户居民作了如下统计，请按户主个人收入进行分类。

表 3.1　某市 2001 年城镇居民户主个人收入数据

X1	职工标准工资收入	X5	单位得到的其他收入
X2	职工奖金收入	X6	其他收入
X3	职工津贴收入	X7	性别
X4	其他工资性收入	X8	就业身份

X1	X2	X3	X4	X5	X6	X7	X8
540.00	0.0	0.0	0.0	0.0	6.00	男	国有
1137.00	125.00	96.00	0.0	109.00	812.00	女	集体
1236.00	300.00	270.00	0.0	102.00	318.00	女	国有
1008.00	0.0	96.00	0.0	86.0	246.00	男	集体
1723.00	419.00	400.00	0.0	122.00	312.00	男	国有
1080.00	569.00	147.00	156.00	210.00	318.00	男	集体
1326.00	0.0	300.00	0.0	148.00	312.00	女	国有
1110.00	110.00	96.00	0.0	80.00	193.00	女	集体
1012.00	88.00	298.00	0.0	79.00	278.00	女	国有
1209.00	102.00	179.00	67.00	198.00	514.00	男	集体
1101.00	215.00	201.00	39.00	146.00	477.00	男	集体

背景

注1：指标选择取决于聚类的目的。

背景

注1：指标选择取决于聚类的目的。

注2：既可用某项指标来分类，也可同时考虑多项指标。显然，分类指标越多，类内元素的共性更多。

背景

注1：指标选择取决于聚类的目的。

注2：既可用某项指标来分类，也可同时考虑多项指标。显然，分类指标越多，类内元素的共性更多。

注3：指标可以是定性或者定量，一般来说有三种尺度，不同尺度的处理方式不大一样。

背景

注1：指标选择取决于聚类的目的。

注2：既可用某项指标来分类，也可同时考虑多项指标。显然，分类指标越多，类内元素的共性更多。

注3：指标可以是定性或者定量，一般来说有三种尺度，不同尺度的处理方式不大一样。

- 间隔尺度：用连续的量来表示。

背景

注1：指标选择取决于聚类的目的。

注2：既可用某项指标来分类，也可同时考虑多项指标。显然，分类指标越多，类内元素的共性更多。

注3：指标可以是定性或者定量，一般来说有三种尺度，不同尺度的处理方式不大一样。

- 间隔尺度：用连续的量来表示。
- 有序尺度：用有序的等级来表示，有次序关系，但没有数量表示。

背景

注1：指标选择取决于聚类的目的。

注2：既可用某项指标来分类，也可同时考虑多项指标。显然，分类指标越多，类内元素的共性更多。

注3：指标可以是定性或者定量，一般来说有三种尺度，不同尺度的处理方式不大一样。

- 间隔尺度：用连续的量来表示。
- 有序尺度：用有序的等级来表示，有次序关系，但没有数量表示。
- 名义尺度：用一些类来表示，既没有等级关系也没有数量关系。

The nominal type differentiates between items or subjects based only on their names or (meta-)categories and other qualitative classifications they belong to; thus dichotomous data involves the construction of classifications as well as the classification of items. Numbers may be used to represent the variables but the numbers do not have numerical value or relationship.

The ordinal type allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted, but still does not allow for relative degree of difference between them.

The interval type allows for the degree of difference between items, but not the ratio between them.

The ratio type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind.

`https://en.wikipedia.org/wiki/Level_of_measurement`

定义

聚类分析，属于一种多元统计分析方法，指根据(多维)样品(或指标)间关系的密切程度进行适当归类，使得同一类中的个体有较大相似性(同质性，Homogeneity)，不同类的个体有较大差异性(异质性，Heterogeneity)。

定义

聚类分析，属于一种多元统计分析方法，指根据(多维)样品(或指标)间关系的密切程度进行适当归类，使得同一类中的个体有较大相似性(同质性，Homogeneity)，不同类的个体有较大差异性(异质性，Heterogeneity)。

注1：在分类之前，对类的个数、类的特征并不清楚。

定义

聚类分析，属于一种多元统计分析方法，指根据(多维)样品(或指标)间关系的密切程度进行适当归类，使得同一类中的个体有较大相似性(同质性，Homogeneity)，不同类的个体有较大差异性(异质性，Heterogeneity)。

注1：在分类之前，对类的个数、类的特征并不清楚。

注2：分类的依据，是样品间相互关系（相似或相近）的密切程度，也是聚类分析的关键。

定义

聚类分析，属于一种多元统计分析方法，指根据(多维)样品(或指标)间关系的密切程度进行适当归类，使得同一类中的个体有较大相似性(同质性，Homogeneity)，不同类的个体有较大差异性(异质性，Heterogeneity)。

注1：在分类之前，对类的个数、类的特征并不清楚。

注2：分类的依据，是样品间相互关系（相似或相近）的密切程度，也是聚类分析的关键。

注3：对样品的分类常称为Q型聚类分析，常用某种距离来刻画；对变量的分类常称为R型聚类分析，常用某种关联指标来度量。

定义

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

`https://en.wikipedia.org/wiki/Cluster_analysis`

第二节：相似性度量

主要内容：常见的距离，常见的相似系数

相似性度量

1. 样品的相似性度量：距离（满足其公理化定义），值越大越相异。
2. 指标的相似性度量：相似系数，值越大越相似。

常见的距离

设$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}, i = 1, 2, \cdots, n,$ 是n 个p 维样品。x_{ij} 表示第i 个样品的第j 个指标, d_{ij} 表示第i 个样品和第j 个样品之间的距离。

常见的距离

设$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$, $i = 1, 2, \cdots, n$, 是n 个p 维样品。x_{ij} 表示第i 个样品的第j 个指标, d_{ij} 表示第i 个样品和第j 个样品之间的距离。

(1) 绝对距离: $d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$

(2) Euclid距离: $d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$

(3) Minkowski距离: $d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q\right]^{\frac{1}{q}}, q > 0$

(4) Chebyshev距离: $d_{ij} = \max_{1 \le k \le p} |x_{ik} - x_{jk}|$

距离

上述距离的的缺点：

- 距离的大小与各指标的观测单位有关，具有一定的随意性；
- 没有考虑指标之间的相关性。

距离

上述距离的的缺点：

- 距离的大小与各指标的观测单位有关，具有一定的随意性；
- 没有考虑指标之间的相关性。

通常的改进办法：

- 当各指标的测量值相差悬殊时，先将数据标准化，然后再计算；
- 马氏距离考虑指标的相关性，且不受量纲影响（线性变换不变性）。

常见的相似系数

(1) 夹角余弦：$C_{ij}=\dfrac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[(\sum_{k=1}^{n}x_{ki}^2)(\sum_{k=1}^{n}x_{kj}^2)\right]^{\frac{1}{2}}}$

(2) 相关系数：$C_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)(x_{kj}-\overline{X}_j)}{\left[\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)^2\sum_{k=1}^{n}(x_{kj}-\overline{X}_j)^2\right]^{\frac{1}{2}}}$

常见的相似系数

(1) 夹角余弦：$C_{ij}=\dfrac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[(\sum_{k=1}^{n}x_{ki}^2)(\sum_{k=1}^{n}x_{kj}^2)\right]^{\frac{1}{2}}}$

(2) 相关系数：$C_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)(x_{kj}-\overline{X}_j)}{\left[\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)^2\sum_{k=1}^{n}(x_{kj}-\overline{X}_j)^2\right]^{\frac{1}{2}}}$

注1：相关系数数即是据标准化后的夹角余弦。

常见的相似系数

(1) 夹角余弦：$C_{ij}=\dfrac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[(\sum_{k=1}^{n}x_{ki}^2)(\sum_{k=1}^{n}x_{kj}^2)\right]^{\frac{1}{2}}}$

(2) 相关系数：$C_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)(x_{kj}-\overline{X}_j)}{\left[\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)^2\sum_{k=1}^{n}(x_{kj}-\overline{X}_j)^2\right]^{\frac{1}{2}}}$

注1：相关系数数即是据标准化后的夹角余弦。

注2：距离和相似系数可以互相转化，如$C_{ij}=1/[1+d_{ij}]$，$d_{ij}=\sqrt{2[1-C_{ij}]}$等。

第三节：类和类的特征

主要内容：类的定义，类的特征，类与类之间的距离，各种类间距离特点对比

类的定义

G表示类，设G中有k个元素，记d_{ij}为$\mathbf{x}_i$和$\mathbf{x}_j$的距离其中$\mathbf{x}_i, \mathbf{x}_j \in G$。

定义1

T为给定阈值，若对于$\forall \mathbf{x}_i, \mathbf{x}_j \in G$，都有$d_{ij} \leq T$，则称$G$为一个类。

类的定义

G表示类，设G中有k个元素，记d_{ij}为$\mathbf{x}_i$和$\mathbf{x}_j$的距离其中$\mathbf{x}_i, \mathbf{x}_j \in G$。

定义1

T为给定阈值，若对于$\forall \mathbf{x}_i, \mathbf{x}_j \in G$，都有$d_{ij} \leq T$，则称$G$为一个类。

定义2

T为给定阈值，若对于$\forall \mathbf{x}_i \in G$，都有$\frac{1}{k-1} \sum_{\mathbf{x}_j \in G} d_{ij} \leq T$，则称$G$为一个类。

类的定义

设类G中有k个元素，记d_{ij}为$\mathbf{x}_i$和$\mathbf{x}_j$的距离，其中$\mathbf{x}_i, \mathbf{x}_j \in G$。

定义3

T和V为给定的阈值，若对于$\forall \mathbf{x}_i, \mathbf{x}_j \in G$，都有$d_{ij} \leq V$，且满足$\dfrac{1}{k(k-1)} \sum_{\mathbf{x}_i \in G} \sum_{\mathbf{x}_j \in G} d_{ij} \leq T$，则称$G$为一个类。

类的定义

设类G中有k个元素，记d_{ij}为$\mathbf{x}_i$和$\mathbf{x}_j$的距离，其中$\mathbf{x}_i, \mathbf{x}_j \in G$。

定义3

T和V为给定的阈值，若对于$\forall \mathbf{x}_i, \mathbf{x}_j \in G$，都有$d_{ij} \leq V$，且满足$\frac{1}{k(k-1)} \sum_{\mathbf{x}_i \in G} \sum_{\mathbf{x}_j \in G} d_{ij} \leq T$，则称$G$为一个类。

定义4

T为给定阈值，若对于$\forall \mathbf{x}_i \in G$，$\exists \mathbf{x}_j \in G$，使得$d_{ij} \leq T$，则称$G$为一个类。

类的定义

设类G中有k个元素，记d_{ij}为$\mathbf{x}_i$和$\mathbf{x}_j$的距离，其中$\mathbf{x}_i, \mathbf{x}_j \in G$。

定义3

T和V为给定的阈值，若对于$\forall \mathbf{x}_i, \mathbf{x}_j \in G$，都有$d_{ij} \leq V$，且满足$\dfrac{1}{k(k-1)} \sum_{\mathbf{x}_i \in G} \sum_{\mathbf{x}_j \in G} d_{ij} \leq T$，则称$G$为一个类。

定义4

T为给定阈值，若对于$\forall \mathbf{x}_i \in G$，$\exists \mathbf{x}_j \in G$，使得$d_{ij} \leq T$，则称$G$为一个类。

注：以上四个定义中，定义1最强，定义2仅次之。

类的特征

记类G中的元素为$\mathbf{x}_i,\ i=1,2,\cdots,m$, 其中$m$为样品数(或指标数)。

(1) 均值(或称重心)：$\bar{\mathbf{x}}_G=\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$

(2) 样本离差阵：$\mathbf{L}_G=\sum_{i=1}^{m}(\mathbf{x}_i-\bar{\mathbf{x}}_G)(\mathbf{x}_i-\bar{\mathbf{x}}_G)^{\mathsf{T}}$

(3) 直径，如$\phi_G=\sum_{i=1}^{m}(\mathbf{x}_i-\bar{\mathbf{x}}_G)^{\mathsf{T}}(\mathbf{x}_i-\bar{\mathbf{x}}_G)=tr(\mathbf{L}_G)$, $\phi_G=\max_{\mathbf{x}_i,\mathbf{x}_j\in G}d_{ij}$

类与类之间的距离

设类G_p与G_q中分别有k个和m个样品，它们的重心分别为$\bar{\mathbf{x}}_p$和$\bar{\mathbf{x}}_q$, 两类之间的距离用$D(p,q)$ 表示。

1. 最短距离法
2. 最长距离法
3. 类平均法
4. 重心法
5. 离差平方和法

类与类之间的距离

最短距离法(nearest neighbor 或single linkage method)

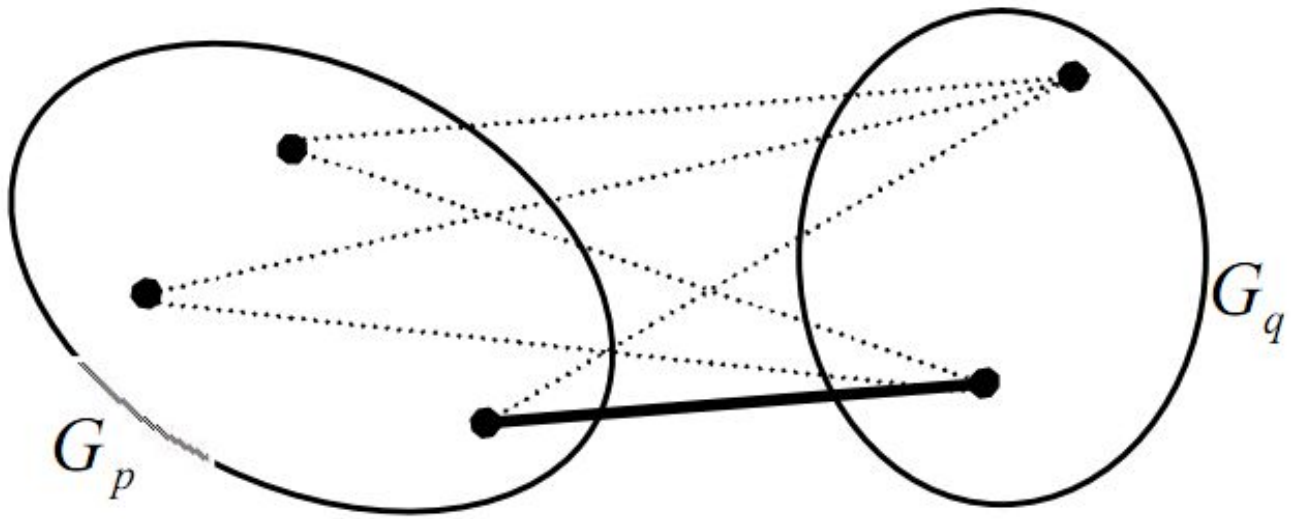
$$D(p,q) = \min_{\mathbf{x}_i \in G_p, \mathbf{x}_j \in G_q} d_{ij}$$

类与类之间的距离

最短距离法(nearest neighbor 或single linkage method)

$$D(p,q) = \min_{\mathbf{x}_i \in G_p, \mathbf{x}_j \in G_q} d_{ij}$$

即定义为两类中最近样品间的距离，强调的是两类样品间的“趋同性”。



类与类之间的距离

最长距离法(farthest neighbor 或complete linkage method)

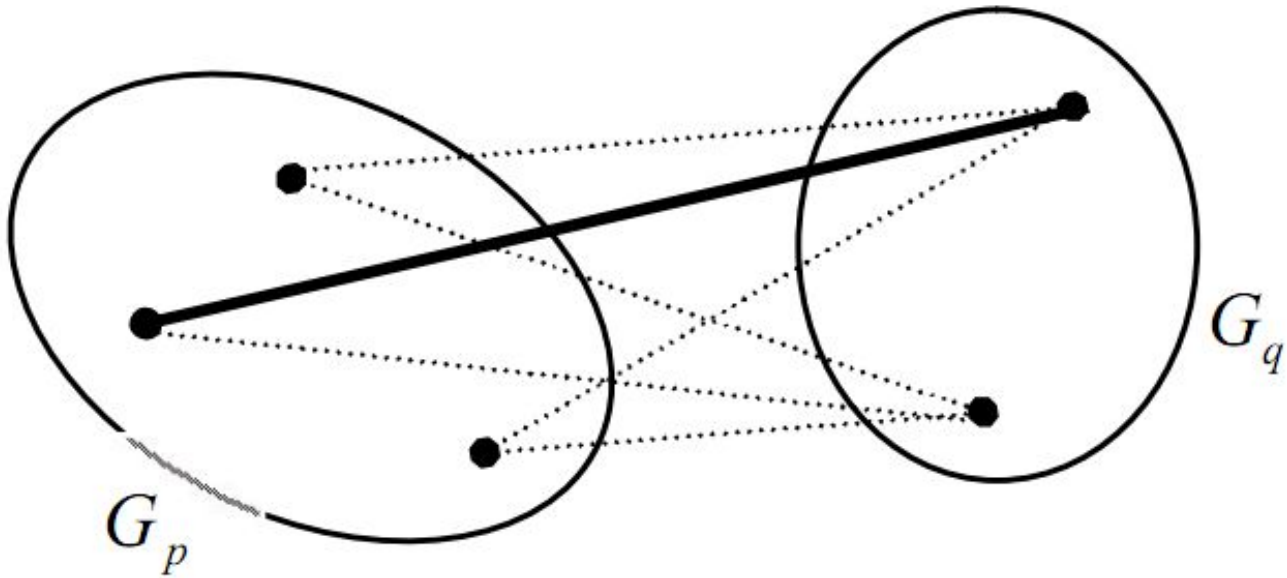
$$D(p,q) = \max_{\mathbf{x}_i \in G_p, \mathbf{x}_j \in G_q} d_{ij}$$

类与类之间的距离

最长距离法(farthest neighbor 或complete linkage method)

$$D(p,q) = \max_{\mathbf{x}_i \in G_p, \mathbf{x}_j \in G_q} d_{ij}$$

即定义为两类间最远样品间的距离，强调的是两类样品间的“趋异性”。



类与类之间的距离

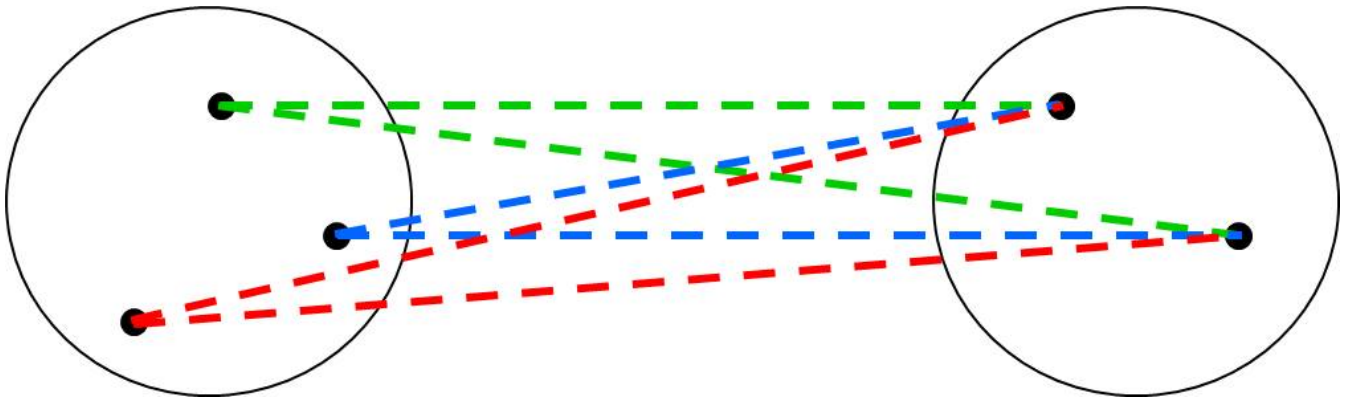
类平均法(group average method)，定义为两类中的样品两两间距离的均值，即

$$D(p,q) = \frac{1}{km} \sum_{\mathbf{x}_i \in G_p} \sum_{\mathbf{x}_j \in G_q} d_{ij}$$

类与类之间的距离

类平均法(group average method)，定义为两类中的样品两两间距离的均值，即

$$D(p,q) = \frac{1}{km} \sum_{\mathbf{x}_i \in G_p} \sum_{\mathbf{x}_j \in G_q} d_{ij}$$



类与类之间的距离

重心法(centroid method)，定义为它们重心间的距离，即。

$$D(p,q) = d_{\bar{\mathbf{x}}_p \bar{\mathbf{x}}_q}$$

类与类之间的距离

离差平方和法(sum of squares method)，借助直径进行定义

$$D(p,q) = \phi_{p+q} - \phi_p - \phi_q$$

其中，ϕ_p，ϕ_q，ϕ_{p+q}分别表示类G_p，G_q和$G_p \cup G_q$的直径。即

$$\phi_{p+q} = \sum_{\mathbf{x}_l \in G_p \cup G_q} (\mathbf{x}_l - \bar{\mathbf{x}})^{\mathsf{T}} (\mathbf{x}_l - \bar{\mathbf{x}})$$

$$\phi_p = \sum_{\mathbf{x}_i \in G_p} (\mathbf{x}_i - \bar{\mathbf{x}}_p)^{\mathsf{T}} (\mathbf{x}_i - \bar{\mathbf{x}}_p)$$

$$\phi_q = \sum_{\mathbf{x}_j \in G_q} (\mathbf{x}_j - \bar{\mathbf{x}}_q)^{\mathsf{T}} (\mathbf{x}_j - \bar{\mathbf{x}}_q)$$

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。
- 最长距离法克服了最短距离法的缺陷，合并后的类与其他类的距离不小于合并前的距离，加大了合并后的类与其他类的距离。

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。
- 最长距离法克服了最短距离法的缺陷，合并后的类与其他类的距离不小于合并前的距离，加大了合并后的类与其他类的距离。
- 重心法虽有很好的代表性，但并未充分利用各样本的信息。

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。
- 最长距离法克服了最短距离法的缺陷，合并后的类与其他类的距离不小于合并前的距离，加大了合并后的类与其他类的距离。
- 重心法虽有很好的代表性，但并未充分利用各样本的信息。
- 类平均法应用较广泛。（1）组间联结法(Between-groups linkage), 只考虑不同类的样品之间距离；（2）组内联结法(Within-groups linkage)，考虑所有样品之间的距离。

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。
- 最长距离法克服了最短距离法的缺陷，合并后的类与其他类的距离不小于合并前的距离，加大了合并后的类与其他类的距离。
- 重心法虽有很好的代表性，但并未充分利用各样本的信息。
- 类平均法应用较广泛。（1）组间联结法(Between-groups linkage), 只考虑不同类的样品之间距离；（2）组内联结法(Within-groups linkage)，考虑所有样品之间的距离。
- 离差平方和法也称Ward法。思想来源于方差分析，即类内离差平方和应当较小，类间离差平方和应当较大。Ward法贪婪求局部最优解：

各种类间距离特点对比

- 最短距离法的主要缺点是链接聚合趋势，实际中不常用。由于类间距离为所有距离中的最短者，则合并后的类与其他类的距离缩小。这样容易形成一个比较大的类，从而大部分样品都被聚在一类中。
- 最长距离法克服了最短距离法的缺陷，合并后的类与其他类的距离不小于合并前的距离，加大了合并后的类与其他类的距离。
- 重心法虽有很好的代表性，但并未充分利用各样本的信息。
- 类平均法应用较广泛。（1）组间联结法(Between-groups linkage), 只考虑不同类的样品之间距离；（2）组内联结法(Within-groups linkage)，考虑所有样品之间的距离。
- 离差平方和法也称Ward法。思想来源于方差分析，即类内离差平方和应当较小，类间离差平方和应当较大。Ward法贪婪求局部最优解：每个样品先自成一类，然后每次选择组内离差平方和增大最小的两类合并，直到所有样品都归为一类为止。

第四节：系统聚类法

主要内容：系统聚类法步骤，方法评述

系统聚类法步骤

系统聚类法(hierarchical clustering method)步骤:

(1) 将n个样品各自看成一类，共有n个类;

(2) 计算类与类间的距离，选择距离最小的两类合并成一个新类，使总类数减少为$n-1$;

(3) 依次类推，每合并一次，减少一类，直至所以样品都合并成一类。

系统聚类法评述

1. 分类的目的不是将全部样品合并成一类，而是通过上述逐渐并类的过程，找到满意的分类方案。

系统聚类法评述

1. 分类的目的不是将全部样品合并成一类，而是通过上述逐渐并类的过程，找到满意的分类方案。
2. 根据聚类的先后以及合并时两类间的距离，画出聚类图(谱系图)，能直观反映样品间的相似程度，从而找到合适的分类方案。

系统聚类法评述

1. 分类的目的不是将全部样品合并成一类，而是通过上述逐渐并类的过程，找到满意的分类方案。
2. 根据聚类的先后以及合并时两类间的距离，画出聚类图(谱系图)，能直观反映样品间的相似程度，从而找到合适的分类方案。
3. 不同的类间距离产生了不同的系统聚类法，合理地定义类间的距离是关键。

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

第二步：将$X5, X6$ 合并为$\{X5, X6\}$ (距离为1);

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

第二步：将$X5, X6$ 合并为$\{X5, X6\}$ (距离为1);

第三步：将$X3, X4$ 合并为$\{X3, X4\}$ (距离为2);

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

第二步：将$X5, X6$ 合并为$\{X5, X6\}$ (距离为1);

第三步：将$X3, X4$ 合并为$\{X3, X4\}$ (距离为2);

第四步：将$\{X3, X4\}, \{X5, X6\}$ 合并为$\{X3, X4, X5, X6\}$ (距离为2);

例3

设有六个样品，每个样品只测了一项指标，它们分别是*1, 2, 5, 7, 9, 10*。样品间的距离取绝对距离，类间距离为最短距离法，将它们归类。

记$X1 = 1, X2 = 2, X3 = 5, X4 = 7, X5 = 9, X6 = 10,$

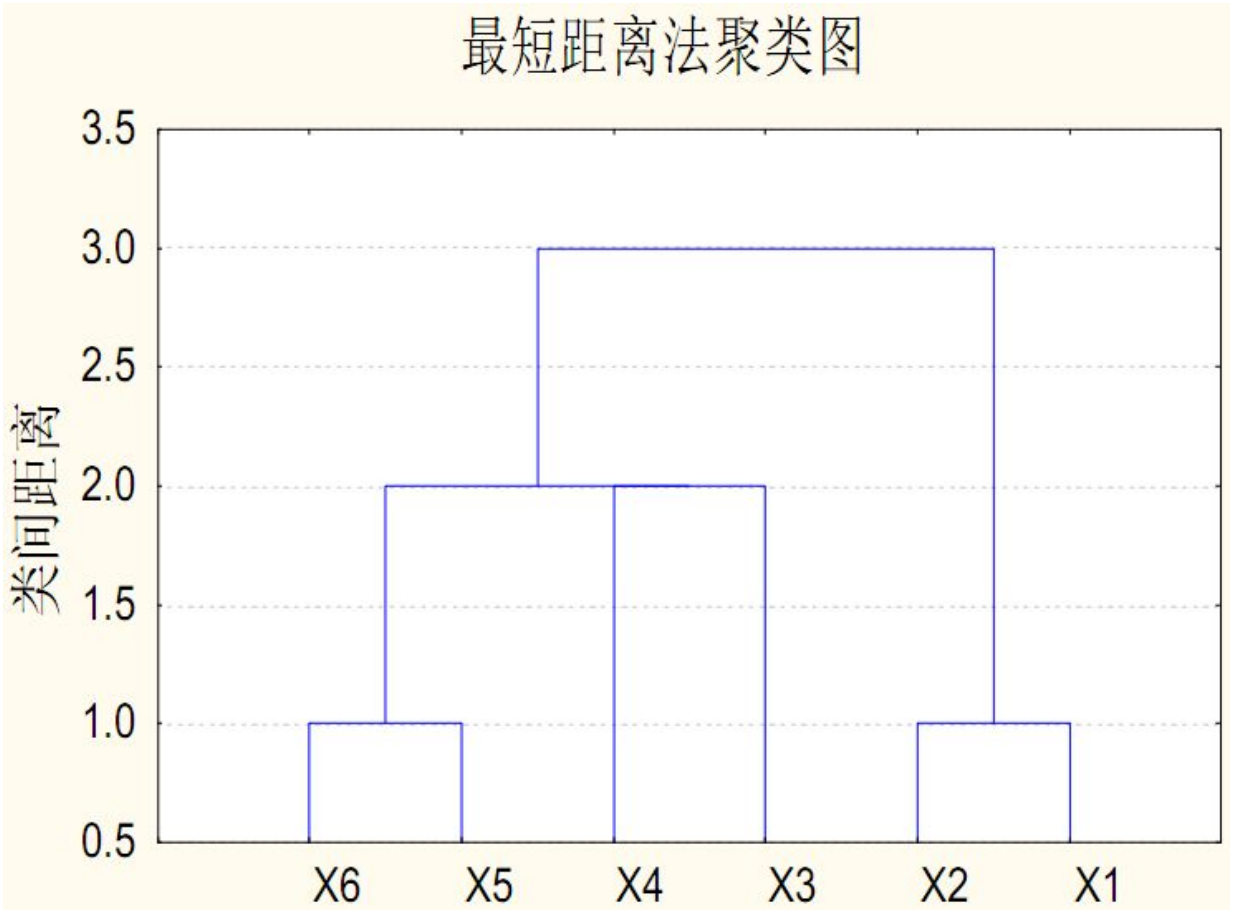
第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

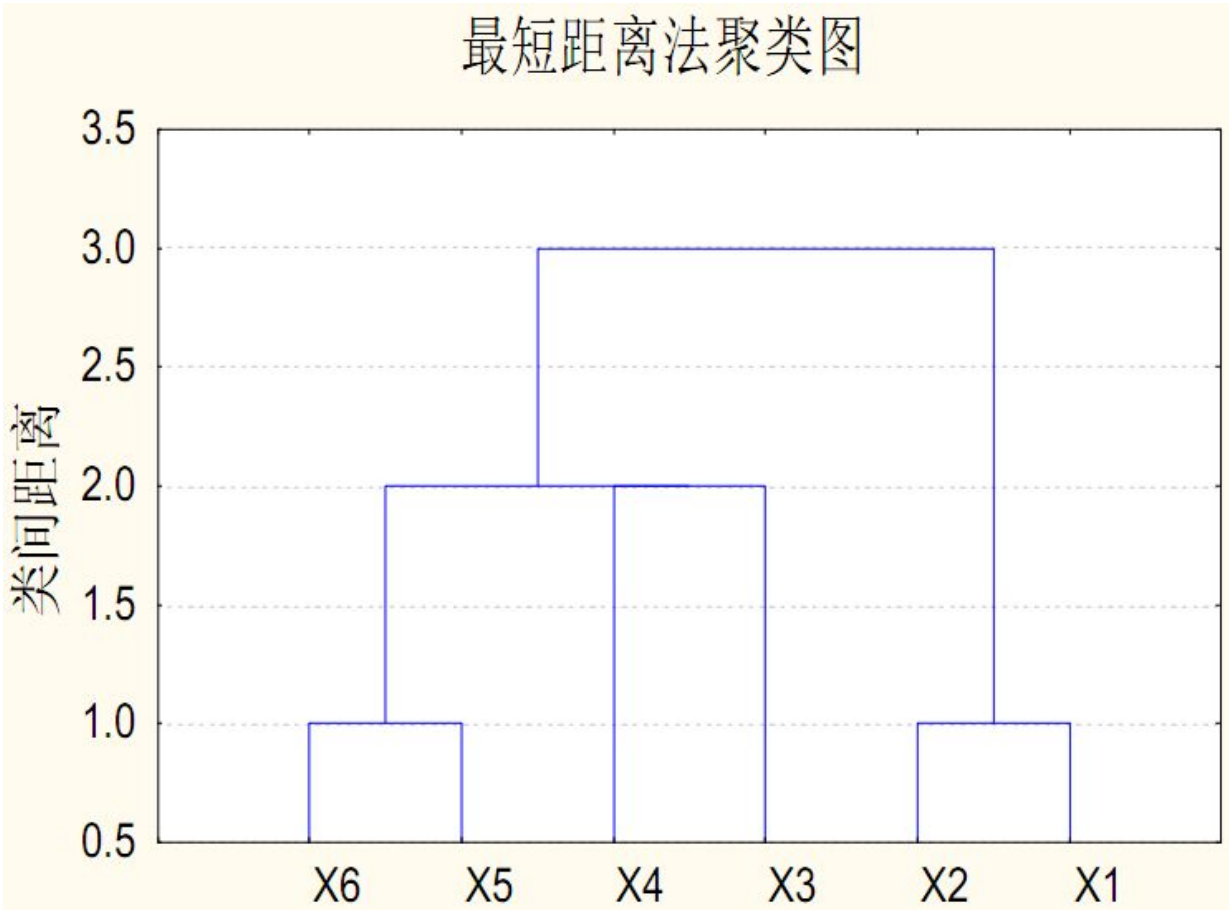
第二步：将$X5, X6$ 合并为$\{X5, X6\}$ (距离为1);

第三步：将$X3, X4$ 合并为$\{X3, X4\}$ (距离为2);

第四步：将$\{X3, X4\}, \{X5, X6\}$ 合并为$\{X3, X4, X5, X6\}$ (距离为2);

第五步：将$\{X1, X2\}, \{X3, X4, X5, X6\}$合并成1类(距离为3), 停止。


最短距离法聚类图
类间距离
3.5
3.0
2.5
2.0
1.5
1.0
0.5
X6
X5
X4
X3
X2
X1



根据聚类图看出，分成两类{1、2}，{5、7、9、10}比较合适。

若定义类间聚类为最长距离。

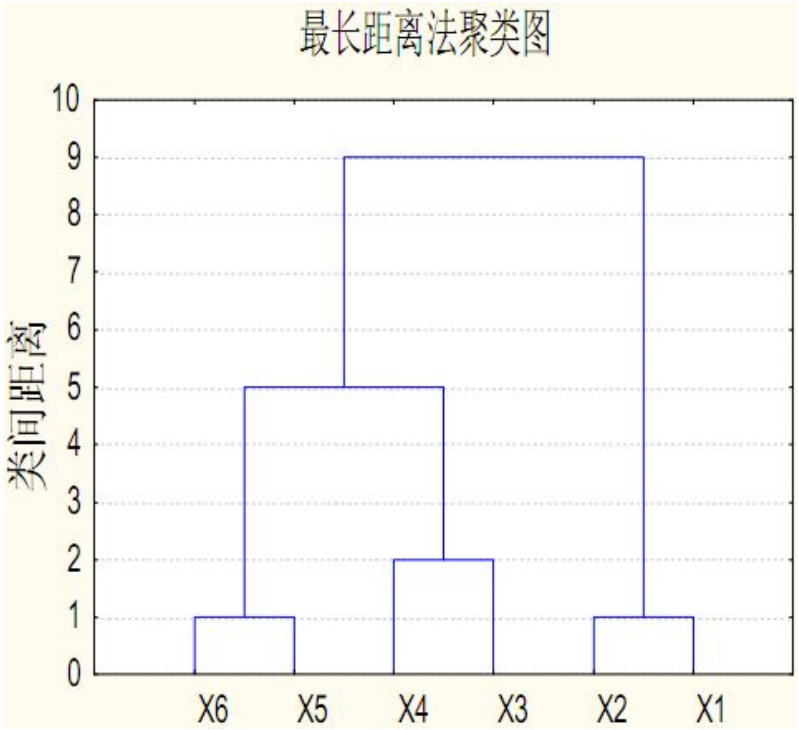
第一步：将$X1, X2$ 合并为$\{X1, X2\}$ (距离为1);

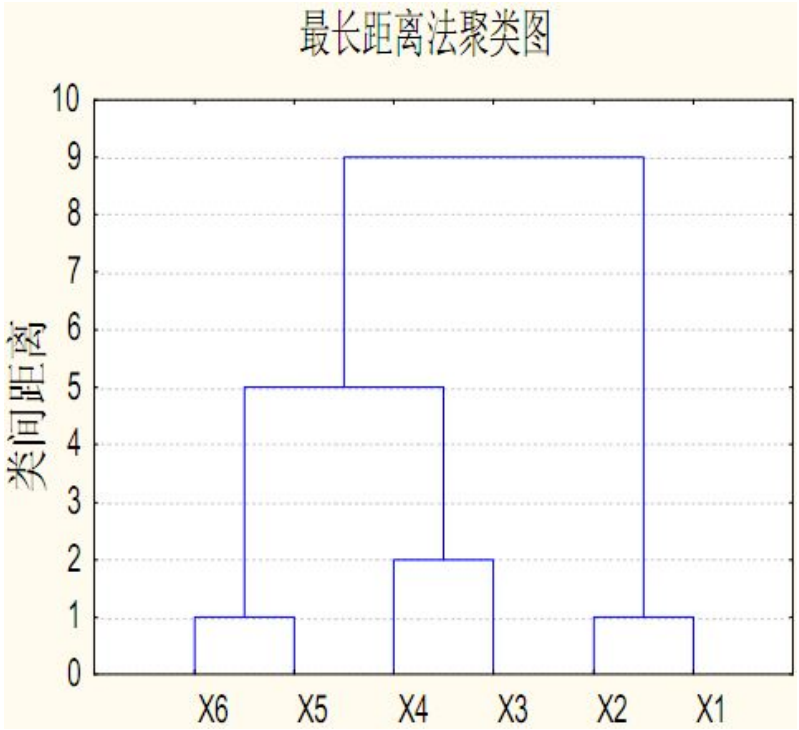
第二步：将$X5, X6$ 合并为$\{X5, X6\}$ (距离为1);

第三步：将$X3, X4$ 合并为$\{X3, X4\}$ (距离为2);

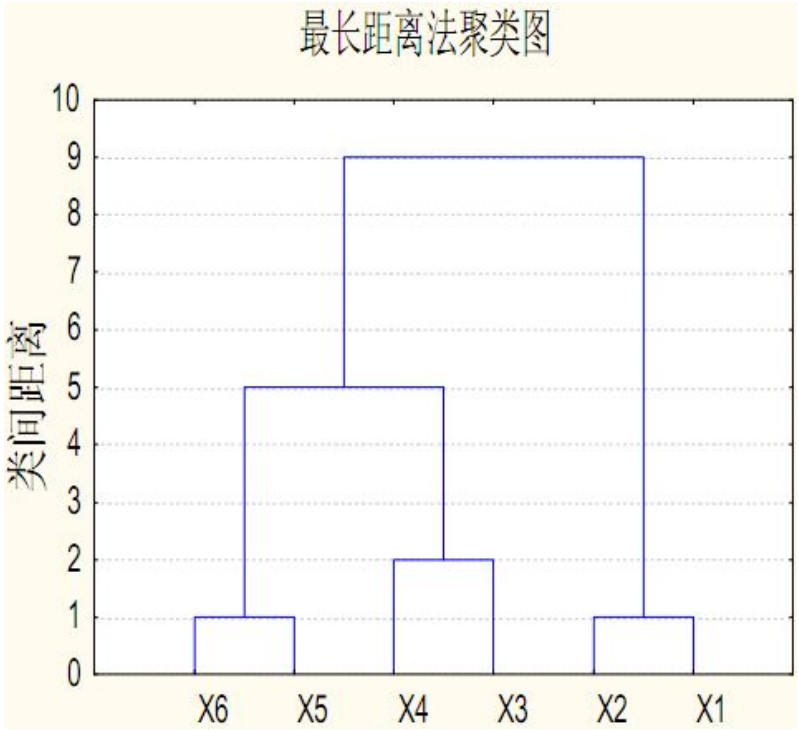
第四步：将$\{X3, X4\}$, $\{X5, X6\}$合并为$\{X3, X4, X5, X6\}$ (距离为5);

第五步：将$\{X1, X2\}$, $\{X3, X4, X5, X6\}$合并为一类(距离为9), 停止。


最长距离法聚类图
类间距离
10
9
8
7
6
5
4
3
2
1
0
X6
X5
X4
X3
X2
X1



若分成两类，{1, 2}, {5, 7, 9, 10}比较合适；若分成三类，{1, 2}, {5, 7}, {9, 10}比较合适。



若分成两类，{1, 2}, {5, 7, 9, 10}比较合适；若分成三类，{1, 2}, {5, 7}, {9, 10}比较合适。

注：不同的方法分类结果不完全一样。究竟采用哪一种，要根据分类问题本身决定。

第五节：K-均值聚类

主要内容：K-均值聚类的步骤和变种

K-均值聚类步骤

K-均值聚类法，也称快速聚类法，由James MacQueen于1967年首次提出。核心思想是，将每个样品聚集到最近的均值所在的类。

K-均值聚类步骤

K-均值聚类法，也称快速聚类法，由James MacQueen于1967年首次提出。核心思想是，将每个样品聚集到最近的均值所在的类。该过程包括以下三步：

1. 给定K个类的均值（如随机），计算每个样品到这些均值的距离，并进行分类；
2. 根据每个类的分类结果，重新计算每个类的均值；
3. 重复以上两步，直至终止规则成立。如固定的迭代步数，各个类的均值之间变化小于阈值，等。

K-均值聚类步骤

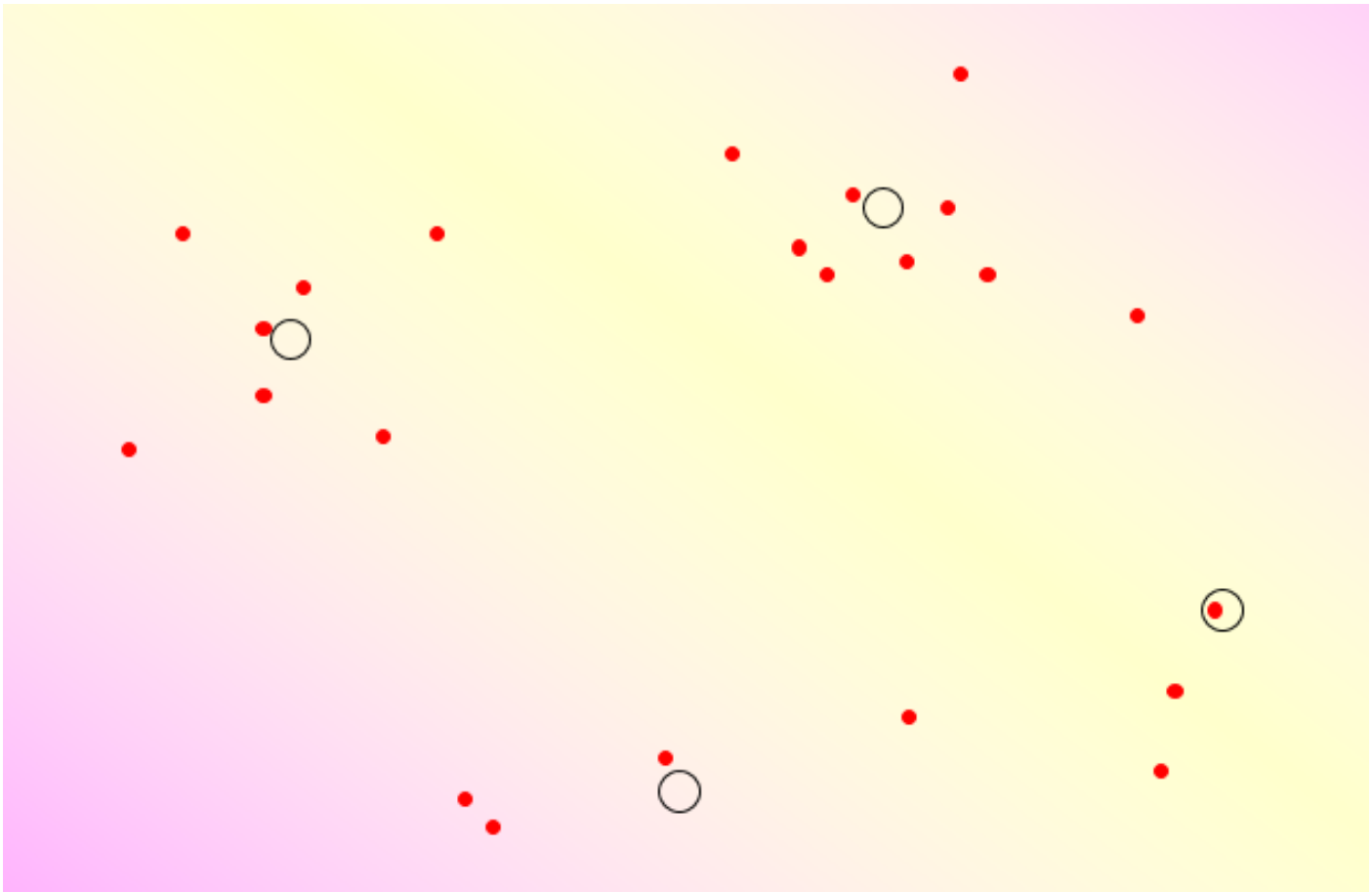
K-均值聚类法，也称快速聚类法，由James MacQueen于1967年首次提出。核心思想是，将每个样品聚集到最近的均值所在的类。该过程包括以下三步：

1. 给定K个类的均值（如随机），计算每个样品到这些均值的距离，并进行分类；
2. 根据每个类的分类结果，重新计算每个类的均值；
3. 重复以上两步，直至终止规则成立。如固定的迭代步数，各个类的均值之间变化小于阈值，等。

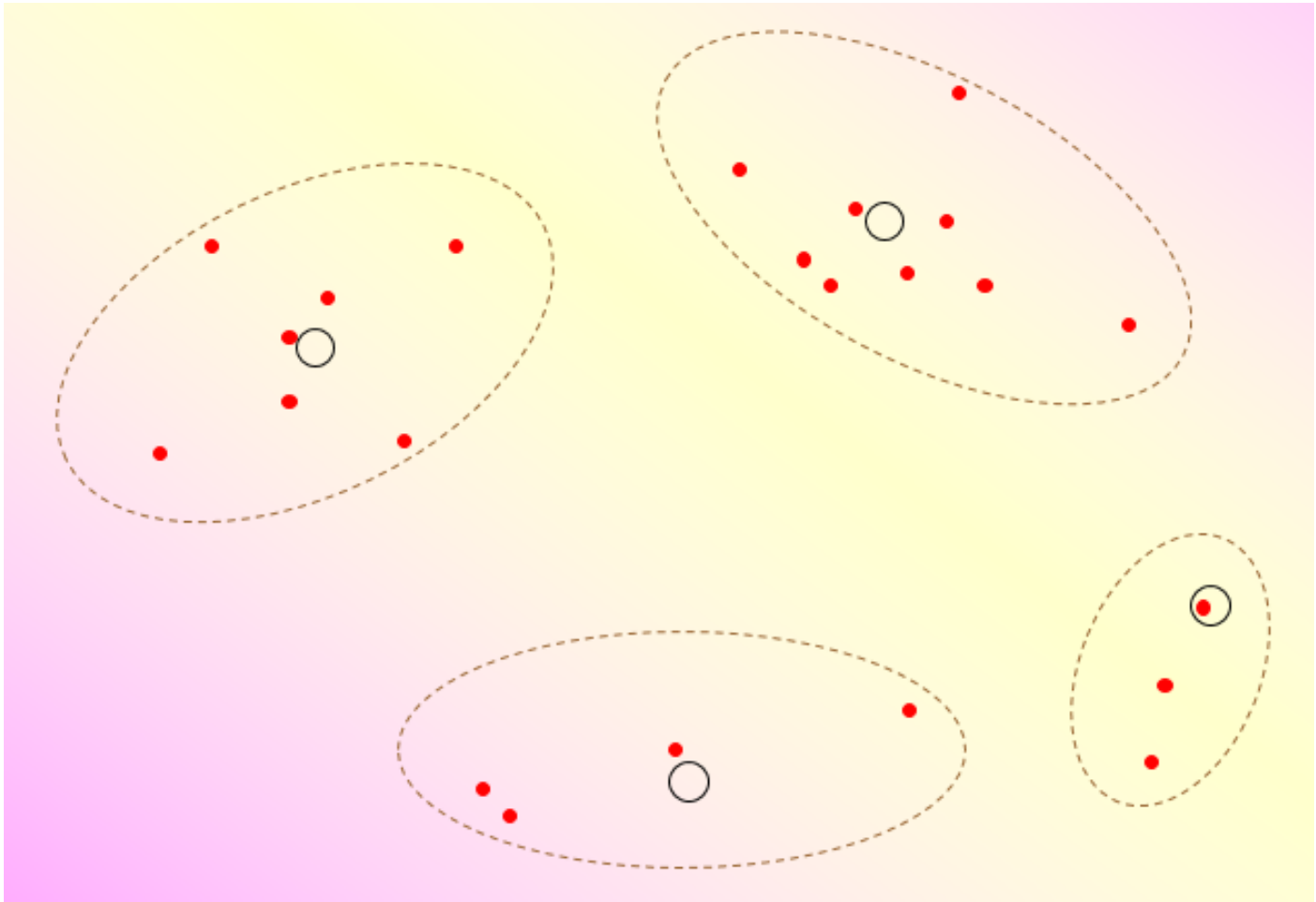
注1：最终的聚类的效果，受到初始均值选择的影响。

注2：聚类效果的稳定性，可用新的初始分类重新运行。

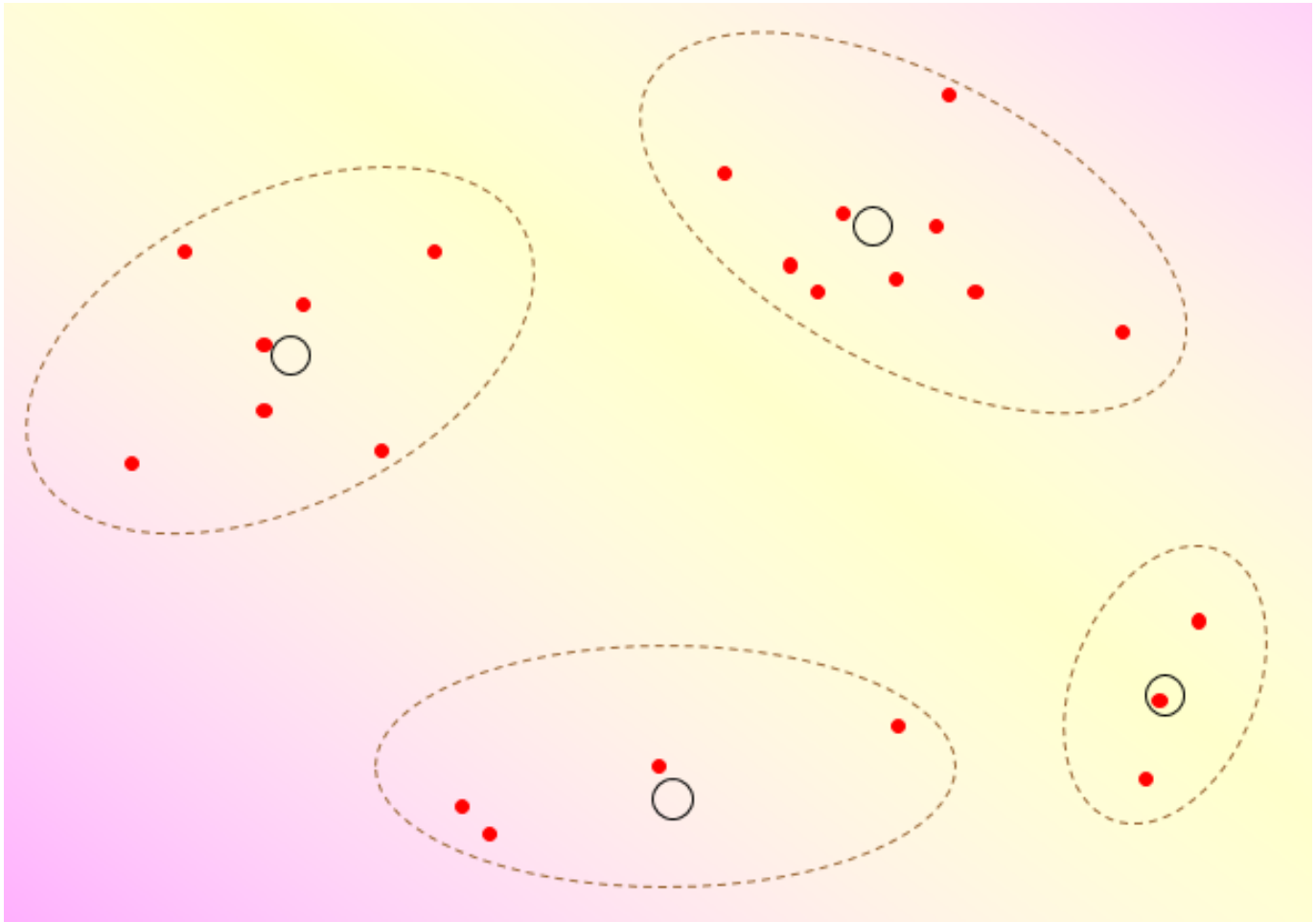
选定初始均值



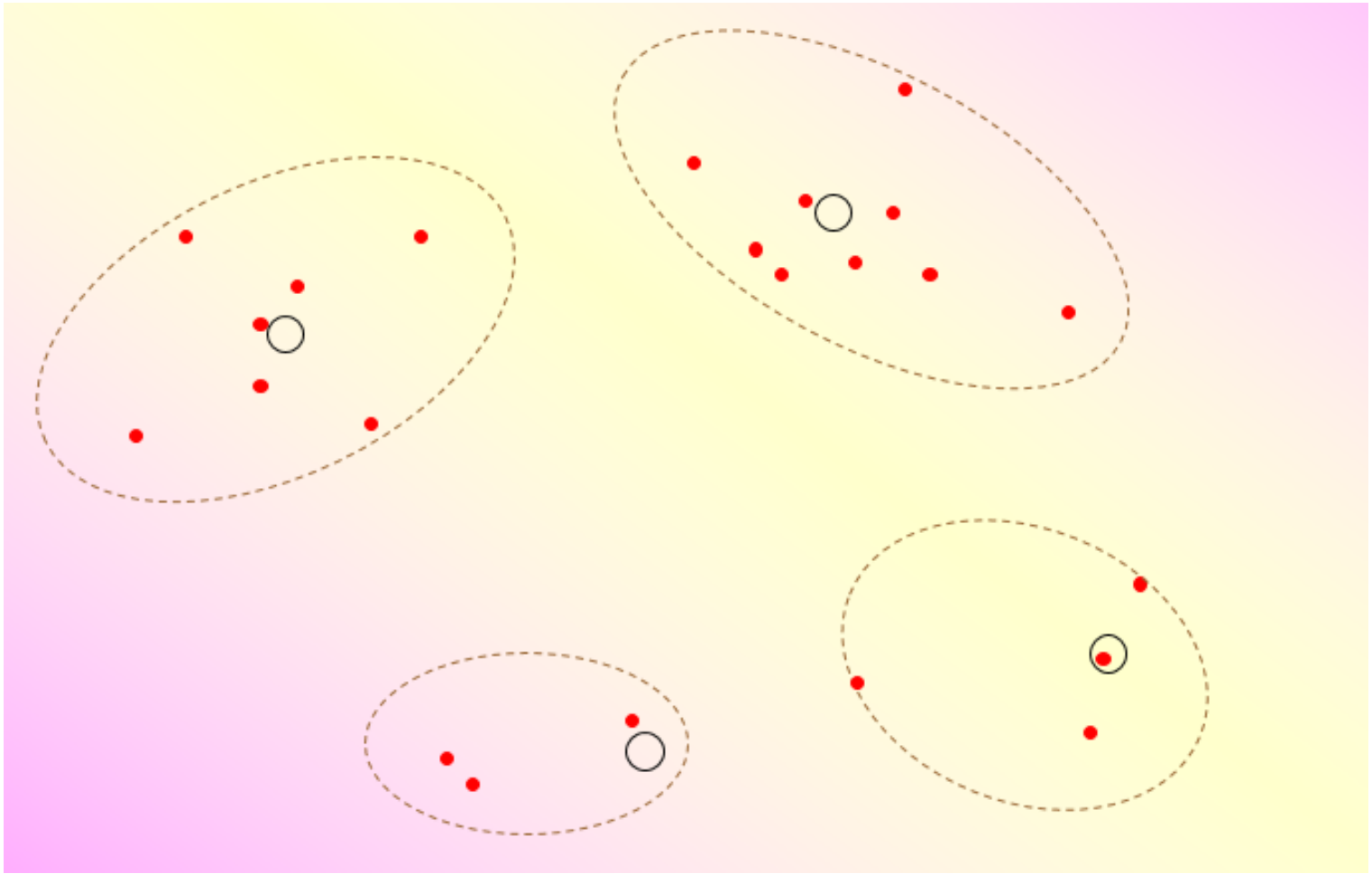
初始聚类



调整均值



调整分类



K-均值法的优缺点

优点：占有内存少、计算量小、处理速度快，特别适合大样本的聚类分析。

缺点：

1. 应用范围有限，要求用户指定分类数目；
2. 只能对样本聚类，不能对变量聚类；
3. 聚类变量必须都是连续性变量。

K均值法变种

K-Modes算法

保留了K-Means的效率，应用于离散数据的聚类。

1. 样本（含多维指标）之间的相似性度量D：若二者的每个分量的属性相同为0，不同则为1，并将所有分量的比较结果相加。因此D越大，不相似程度越强。
2. 更新每个类的modes，分别用每个指标出现频率最大的那个属性值作为代表。

K均值法变种

K-Prototype算法

结合K-Means与K-Modes算法，应用于离散与连续混合时的聚类。

1. 样本之间的相似性度量D=P1+a*P2，其中P1为连续变量采用K-means方法计算相似性程程度；P2为离散变量采用K-modes方法计算相似性程度；a是离散变量的权重。
2. 更新每个类的prototype，分别针对每个变量进行，若为离散变量采用最大频率，连续变量采用平均值。

第六节：分类数的确定

如何选择分类数，是各种聚类方法都需要面临的问题。如分类数是K-Means聚类法的输入参数，系统聚类法最终得到的只是一个树状结构图，仍存在很多分类方案。

第六节：分类数的确定

如何选择分类数，是各种聚类方法都需要面临的问题。如分类数是K-Means聚类法的输入参数，系统聚类法最终得到的只是一个树状结构图，仍存在很多分类方案。

确定分类数的主要障碍，是对类的结构和特征很难给出统一定义，这样就给不出从理论上和实践中都可行的虚无假设。实际应用中，人们主要根据研究的目的选择合适的分类数。

分类数的确定

Demirmen曾提出了根据树状结构图分类的准则：

1. 各类重心之间的距离尽量大;
2. 各类所包含的元素都不要过分多;
3. 分类数应该符合使用的目的;
4. 不同聚类方法发现的类尽量相同。

分类数的确定

1. 给定类与类之间距离的阈值。事实：当聚类过程中，每次合并的两类的类间距离单调增加。达到该阈值的分类数作为合适的分类数。

分类数的确定

1. 给定类与类之间距离的阈值。事实：当聚类过程中，每次合并的两类的类间距离单调增加。达到该阈值的分类数作为合适的分类数。
2. 系统聚类中每次合并时的类间的距离，称之为聚合系数。聚合系数随分类数的变化曲线，变化平缓之处作为合适的分类数。

分类数的确定

1. 给定类与类之间距离的阈值。事实：当聚类过程中，每次合并的两类的类间距离单调增加。达到该阈值的分类数作为合适的分类数。
2. 系统聚类中每次合并时的类间的距离，称之为聚合系数。聚合系数随分类数的变化曲线，变化平缓之处作为合适的分类数。
3. 考虑类间离差平方和SSA与总离差平方和SST的比值$R^2 = \frac{SSA}{SST}$（统计量）。事实：分类越多，类内离差平方和越小，则R^2越大。R^2随分类数变化的曲线，变化平缓之处作为合适的分类数。

分类数的确定

1. 给定类与类之间距离的阈值。事实：当聚类过程中，每次合并的两类的类间距离单调增加。达到该阈值的分类数作为合适的分类数。
2. 系统聚类中每次合并时的类间的距离，称之为聚合系数。聚合系数随分类数的变化曲线，变化平缓之处作为合适的分类数。
3. 考虑类间离差平方和SSA与总离差平方和SST的比值$R^2 = \frac{SSA}{SST}$（统计量）。事实：分类越多，类内离差平方和越小，则R^2越大。R^2随分类数变化的曲线，变化平缓之处作为合适的分类数。
4. 类似R^2，伪统计量$F = \frac{SSA/r-1}{SSE/n-r}$，其中为$n$样本数，$r$为类别数。

第七节：有序样本聚类

主要内容：最优分割法步骤


上证指数(000001)
沪A 上涨：814家 平盘：63家 下跌：468家
沪B 上涨：16家 平盘：25家 下跌
深A 上涨：985家 平盘：106家 下跌：850家
深B 上涨：31家 平盘：4家 下跌：
2934.47
18.74
0.64%
2018-06-21 10:48:28
今 开：2912.00
最 高：2940.59
成交量：656935手
昨 收：2915.73
最 低：2908.01
成交额：6733344万元
指数行情
行情走势
每日行情
每周行情
每月行情
股票排行
行业板块
概念板块
操盘必读
分时
五日
一年
日K
周K
月K
5分
15分
30分
60分
2018/06/20 10:02 价位:2903.33 均价:- 成交量:4666.76 涨跌幅:-0.15%
3228.25
3183.64
3139.03
3094.41
3049.80
3005.19
2960.58
2915.96
2871.35
5.85%
4.39%
2.93%
1.46%
0.00%
-1.46%
-2.93%
-4.39%
-5.85%
2018/06/14
2018/06/15
2018/06/19
2018/06/20
2018/06/21

前几节叙述的方法中，各样品彼此平等，相互独立。如果要求样品按照一定顺序（时间，空间等）排列，且分类时不能打乱顺序，则属于有序样品聚类。

前几节叙述的方法中，各样品彼此平等，相互独立。如果要求样品按照一定顺序（时间，空间等）排列，且分类时不能打乱顺序，则属于有序样品聚类。

有序样品的分类，实质上是寻找一些分割点，将有序样本划分为几个分段，每个分段看成一类。不同的分割点得到不同的分类，最优的分割是分段内的差异尽量小，分段之间的差异尽量大。

前几节叙述的方法中，各样品彼此平等，相互独立。如果要求样品按照一定顺序（时间，空间等）排列，且分类时不能打乱顺序，则属于有序样品聚类。

有序样品的分类，实质上是寻找一些分割点，将有序样本划分为几个分段，每个分段看成一类。不同的分割点得到不同的分类，最优的分割是分段内的差异尽量小，分段之间的差异尽量大。

1958年，Fisher基于离差平方和提出有序聚类的最优分割方法。

最优分割法

设有序样本为$\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \cdots, \mathbf{X}_{(n)}$，每个为$m$维向量。用$b(n,k)$表示将$n$个有序样品分为$k$类的某种分法，记为

- $G_1 = \{i_1, i_1+1, \cdots, i_2-1\}$,
- $G_2 = \{i_2, i_2+1, \cdots, i_3-1\}$,
- $\cdots$,
- $G_k = \{i_k, i_k+1, \cdots, i_{k+1}-1\}$

其中分点为$1 = i_1 < i_2 < \cdots < i_k < i_{k+1} - 1 = n$

最优分割法

设某一类包含的样本为$\mathbf{X}_{(i)}, \mathbf{X}_{(i+1)}, \cdots, \mathbf{X}_{(j)}$，其中$i < j$，记为$G = \{i, i+1, \cdots, j\}$。该类的直径$D(i,j)$定义为

$$D(i,j) = \sum_{t=i}^{j} (\mathbf{X}_{(t)} - \overline{\mathbf{X}}_G)^{\mathsf{T}} (\mathbf{X}_{(t)} - \overline{\mathbf{X}}_G)$$

其中$\overline{\mathbf{X}}_G$为该类的均值（向量），即$\overline{\mathbf{X}}_G = \frac{1}{j-i+1} \sum_{t=i}^{j} \mathbf{X}_{(t)}$。

最优分割法

设某一类包含的样本为$\mathbf{X}_{(i)}, \mathbf{X}_{(i+1)}, \cdots, \mathbf{X}_{(j)}$，其中$i < j$，记为$G = \{i, i+1, \cdots, j\}$。该类的直径$D(i,j)$定义为

$$D(i,j) = \sum_{t=i}^{j} (\mathbf{X}_{(t)} - \overline{\mathbf{X}}_G)^{\mathsf{T}} (\mathbf{X}_{(t)} - \overline{\mathbf{X}}_G)$$

其中$\overline{\mathbf{X}}_G$为该类的均值（向量），即$\overline{\mathbf{X}}_G = \frac{1}{j-i+1} \sum_{t=i}^{j} \mathbf{X}_{(t)}$。还可以定义$\overline{\mathbf{X}}_G$为该类的中位数。特别地，当$m = 1$时

$$D(i,j) = \sum_{t=i}^{j} (\mathbf{X}_{(t)} - \overline{\mathbf{X}}_G)^2$$

最优分割法

设有序样本某种分类$b(n,k)$的损失函数定义为

$$L[b(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1} - 1)$$

对于固定的n,k，损失函数$L[b(n,k)]$越小表示各类的离差平方和越小，分类越合理。寻求最小损失对应的分类，记为$P(n,k)$。

最优分割法

Fisher最优分类的递推公式

$$L[b(n,k)] = \min_{k \le j \le n} \{L[P(j-1,k-1)] + D(j,n)\}, k > 2$$

当$k = 2$时，$L[b(n,2)] = \min_{2 \le j \le n}\{D(i,j-1) + D(j,n)\}$。

最优分割法

Fisher最优分类的递推公式

$$L[b(n,k)] = \min_{k \le j \le n} \{L[P(j-1,k-1)] + D(j,n)\}, k > 2$$

当$k = 2$时，$L[b(n,2)] = \min_{2 \le j \le n}\{D(i,j-1) + D(j,n)\}$。

递推公式表明，将n个样本划分为k类的最优分割，应建立在将$j-1$个样本划分为$k-1$类的最优分割基础之上，其中$j = 2,3,\cdots,n$。

最优分割法

最优分类的求法：

首先，找到分点j_k使得递推公式达到极小，即

$$L[P(n,k)] = L[P(j_k - 1, k-1)] + D(j_k, n)$$

从而得到第k类，$G_k = \{j_k, j_k + 1, \cdots, n\}$。

最优分割法

最优分类的求法：

首先，找到分点j_k使得递推公式达到极小，即

$$L[P(n,k)] = L[P(j_k - 1, k-1)] + D(j_k, n)$$

从而得到第k类，$G_k = \{j_k, j_k + 1, \cdots, n\}$。

然后寻找j_{k-1}满足

$$L[P(j_k - 1, k-1)] = L[P(j_{k-1} - 1, k-2)] + D(j_{k-1}, j_k - 1)$$

从而得到第$k-1$类，$G_{k-1} = \{j_{k-1}, j_{k-1} + 1, \cdots, j_k - 1\}$。

最优分割法

最优分类的求法：

首先，找到分点j_k使得递推公式达到极小，即

$$L[P(n,k)] = L[P(j_k - 1, k-1)] + D(j_k, n)$$

从而得到第k类，$G_k = \{j_k, j_k + 1, \cdots, n\}$。

然后寻找j_{k-1}满足

$$L[P(j_k - 1, k-1)] = L[P(j_{k-1} - 1, k-2)] + D(j_{k-1}, j_k - 1)$$

从而得到第$k-1$类，$G_{k-1} = \{j_{k-1}, j_{k-1} + 1, \cdots, j_k - 1\}$。

依次类推求得其他类，即有最优分割$P(n,k) = \{G_1, G_2, \cdots, G_k\}$。

最优分割法

例4

为了了解儿童生长发育规律，统计了男孩*1-11*岁每年的平均增加重量*(kg)*，如下表所示：

年龄	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
增加量	*9.3*	*1.8*	*1.9*	*1.7*	*1.5*	*1.3*	*1.4*	*2.0*	*1.9*	*2.3*	*2.1*

试问：可分为哪几个发育阶段？

最优分割法

第一步，计算直径$D(i,j), 1 \le i \le j \le n$。

j \ i	1	2	3	4	5	6	7	8	9	10
2	28.125									
3	37.007	0.005								
4	42.208	0.020	0.020							
5	45.992	0.088	0.080	0.020						
6	49.128	0.232	0.200	0.080	0.020					
7	51.100	0.280	0.232	0.088	0.020	0.005				
8	51.529	0.417	0.393	0.308	0.290	0.287	0.180			
9	51.980	0.469	0.454	0.393	0.388	0.370	0.207	0.005		
10	52.029	0.802	0.800	0.774	0.773	0.708	0.420	0.087	0.080	
11	52.182	0.909	0.909	0.895	0.889	0.793	0.452	0.088	0.080	0.020

例如，类$G = X_{(5)}, X_{(6)}, X_{(7)}$的直径：$\overline{\mathbf{X}}_G = \frac{1}{3}(1.5 + 1.3 + 1.4) = 1.4,$

$D(5,7) = (1.5 - 1.4)^2 + (1.3 - 1.4)^2 + (1.4 - 1.4)^2 = 0.02$

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots$,时，最优分割的损失。

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots$,时，最优分割的损失。

首先计算$L[P(l,2)], 3 \leq l \leq 11$。

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots$,时，最优分割的损失。

首先计算$L[P(l,2)], 3 \leq l \leq 11$。例如当$l = 3$时，若将前三个样本划分成2类，存在两种可能结果$\{1\}, \{2,3\}$，或$\{1,2\}, \{3\}$。最小损失为

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots,$时，最优分割的损失。

首先计算$L[P(l,2)], 3 \leq l \leq 11$。例如当$l = 3$时，若将前三个样本划分成2类，存在两种可能结果$\{1\},\{2,3\}$，或$\{1,2\},\{3\}$。最小损失为

$$\begin{aligned} L[P(3,2)] &= \min\{D(1,1) + D(2,3), D(1,2) + D(3,3)\} \\ &= \min\{0 + 0.005, 28.125 + 0\} \\ &= 0.005 \end{aligned}$$

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots$,时，最优分割的损失。

其次计算$L[P(l,3)], 4 \leq l \leq 11$。

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots,$时，最优分割的损失。

其次计算$L[P(l,3)], 4 \leq l \leq 11$。例如当$l=4$时，若将前四个样本划分成3类，根据递推公式，最小损失为

$$\begin{aligned} L[P(4,3)] &= \min\{L(P(2,2)) + D(3,4), L(P(3,2)) + D(4,4)\} \\ &= \min\{0 + 0.020, 0.005 + 0\} \\ &= 0.005 \end{aligned}$$

最优分割法

第二步，计算最小损失函数$L[P(l,k)], 2 \leq k \leq 10, k+1 \leq l \leq 11$。即分别计算将$l$个样本分成$2, 3, \cdots$,时，最优分割的损失。

依次类推，计算$L[P(l,4)], L[P(l,5)], \cdots, L[P(l,10)]$。如下表，其中括号内的数字是最优分割点的位置。

最优分割法

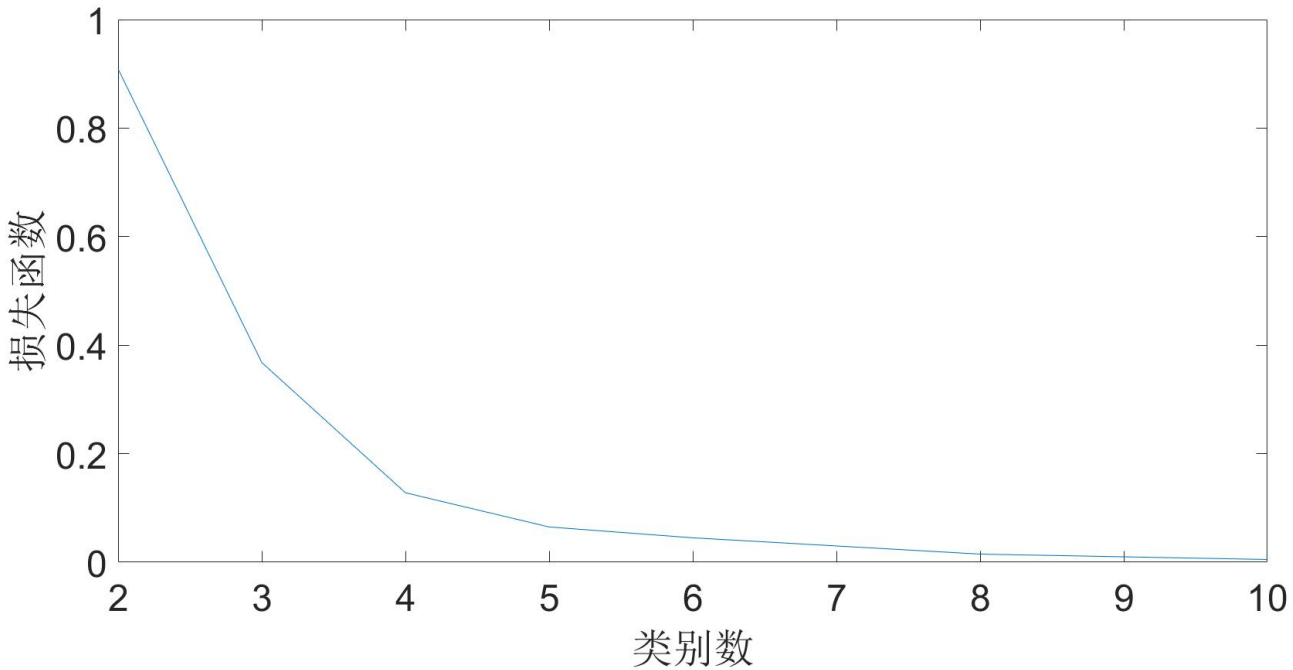
表: 最优损失函数

l \ k	2	3	4	5	6	7	8	9	1
3	0.005 (2)								
4	0.020 (2)	0.005 (4)							
5	0.088 (2)	0.020 (5)	0.005 (5)						
6	0.232 (2)	0.040 (5)	0.020 (6)	0.005 (6)					
7	0.280 (2)	0.040 (5)	0.025 (6)	0.010 (6)	0.005 (6)				
8	0.417 (2)	0.280 (8)	0.040 (8)	0.025 (8)	0.010 (8)	0.005 (8)			
9	0.469 (2)	0.285 (8)	0.045 (8)	0.030 (8)	0.015 (8)	0.010 (3)	0.005 (8)		
10	0.802 (2)	0.367 (8)	0.127 (8)	0.045 (10)	0.030 (10)	0.015 (10)	0.010 (10)	0.005 (8)	
11	0.909 (2)	0.368 (8)	0.128 (8)	0.065 (10)	0.045 (11)	0.030 (11)	0.015 (11)	0.010 (11)	0.005

最优分割法

确定类别数。

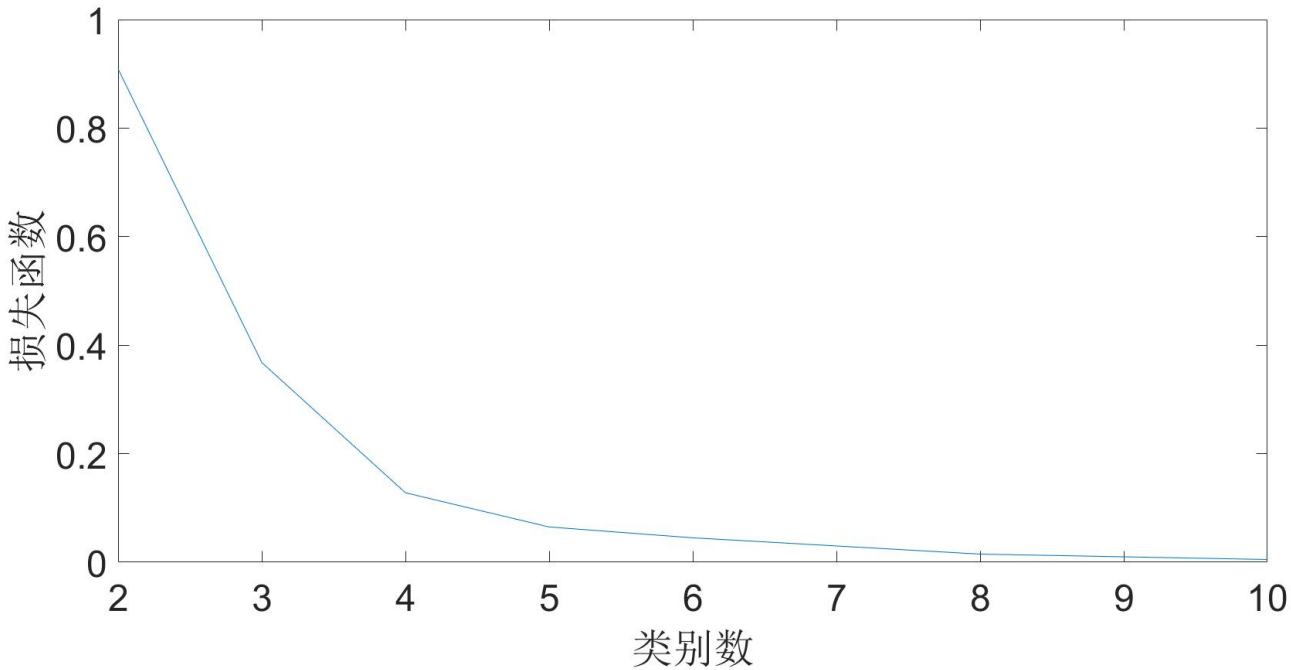
最优损失函数$L[P(11,k)]$关于类别数k的变化曲线



最优分割法

确定类别数。

最优损失函数$L[P(11,k)]$关于类别数k的变化曲线



图形表明，$k=3,4$比较合理。

最优分割法

l \ k	2	3	4	5	6	7	8	9	1
3	0.005 (2)								
4	0.020 (2)	0.005 (4)							
5	0.088 (2)	0.020 (5)	0.005 (5)						
6	0.232 (2)	0.040 (5)	0.020 (6)	0.005 (6)					
7	0.280 (2)	0.040 (5)	0.025 (6)	0.010 (6)	0.005 (6)				
8	0.417 (2)	0.280 (8)	0.040 (8)	0.025 (8)	0.010 (8)	0.005 (8)			
9	0.469 (2)	0.285 (8)	0.045 (8)	0.030 (8)	0.015 (8)	0.010 (3)	0.005 (8)		
10	0.802 (2)	0.367 (8)	0.127 (8)	0.045 (10)	0.030 (10)	0.015 (10)	0.010 (10)	0.005 (8)	
11	0.909 (2)	0.368 (8)	0.128 (8)	0.065 (10)	0.045 (11)	0.030 (11)	0.015 (11)	0.010 (11)	0.005

分成3类：$G_1 = \{X_{(1)}\}; G_2 = \{X_{(2)}, \cdots, X_{(7)}\}; G_3 = \{X_{(8)}, \cdots, X_{(11)}\}$

最优分割法

l \ k	2	3	4	5	6	7	8	9	1
3	0.005 (2)								
4	0.020 (2)	0.005 (4)							
5	0.088 (2)	0.020 (5)	0.005 (5)						
6	0.232 (2)	0.040 (5)	0.020 (6)	0.005 (6)					
7	0.280 (2)	0.040 (5)	0.025 (6)	0.010 (6)	0.005 (6)				
8	0.417 (2)	0.280 (8)	0.040 (8)	0.025 (8)	0.010 (8)	0.005 (8)			
9	0.469 (2)	0.285 (8)	0.045 (8)	0.030 (8)	0.015 (8)	0.010 (3)	0.005 (8)		
10	0.802 (2)	0.367 (8)	0.127 (8)	0.045 (10)	0.030 (10)	0.015 (10)	0.010 (10)	0.005 (8)	
11	0.909 (2)	0.368 (8)	0.128 (8)	0.065 (10)	0.045 (11)	0.030 (11)	0.015 (11)	0.010 (11)	0.00

分成4类：$G_1 = \{X_{(1)}\}; G_2 = \{X_{(2)}, X_{(3)}, X_{(4)}\}; G_3 = \{X_{(5)}, X_{(6)}, X_{(7)}\};$

$G_4 = \{X_{(8)}, \cdots, X_{(11)}\}$

延伸阅读

1. Leonard Kaufman, Peter J. Rousseeuw，Finding groups in data: an introduction to cluster analysis，Wiley &Sons，2005
2. 方开泰，聚类分析，北京：地质出版社, 1982.

1. 根据CH3 exercise.xls的数据，操作SPSS软件或其他软件，采用系统聚类法和K均值法进行聚类，聚类数目请自行设定，合理为宜。
2. 编程实现有序样品聚类的最优分割法。收集某只股票的数据，或我国商品房销售价格数据，或其他经济时间序列数据，进行实证分析。

要求：（1）分组完成，每组最多5人；（2）作业1须有结果分析，如图表有文字解释；（3）作业2需提供数据出处，原始数据，程序代码，结果分析；（4）提交纸质版，封面为组内成员的姓名和学号。