模式识别

第2讲 聚类分析

2018~2019学年

内容安排

一、绪论、数学基础（第1讲）

二、聚类分析（第2讲）

三、判别函数分类法（几何分类法）（第3、4讲）

四、统计决策分类法（概率分类法）（第5、6讲）

五、特征提取与选择（第7讲）

六、模糊模式识别（第8讲）

七、神经网络模式识别（第9讲）

期末考试（平时作业：40%，期末考试：60%）

二、聚类分析

2.1　相似性聚类的概念

2.2　相似性测度和聚类准则

2.3　基于距离阈值的聚类算法

2.4　层次聚类法

2.5　动态聚类法

2.6　聚类结果评价

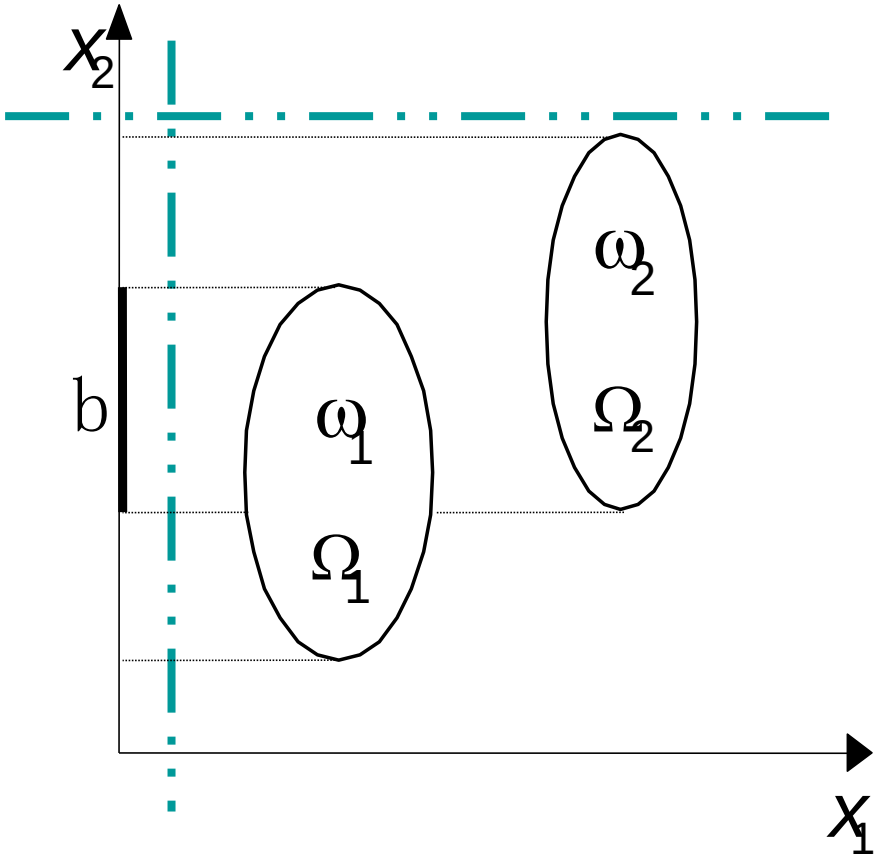
2.1 相似性聚类的概念

- **无监督学习**：使用**不知类别**的样本集进行分类器设计
  - 基于概率密度函数估计的方法（不讲）
  - 基于样本间相似性度量的方法（**聚类分析**）
- **聚类分析**：是指在没有太多先验知识的情况下，按“**物以类聚**”思想，根据模式间的**某种相似性**，对样本进行分类。因此也称为**相似性聚类**
  - 训练前，甚至没有确切的类别数目和类别定义，需要根据待分类样本集的实际特征分布情况与分类活动的应用目的，通过训练样本来学习出类别数目和“类别的操作定义”，同时为训练样本分配类别
  - 一般需要迭代多次才会得出有意义的结果
  - “物以类聚”原则展开说就是：
    “**同类样本间的相似性 大于 不同类样本间的相似性**”。
  - 聚类方法的有效性：取决于**分类算法**与样本**特征分布**的**匹配**
  - 从方法丰富性与成熟度看，与判别分析（分类）方法相比，聚类方法更多利用直观思想和启发式方法，方法较多，但是缺乏完整稳定的聚类理论基础

特征设计对聚类分析的影响

（1）特征选取不当使分类无效

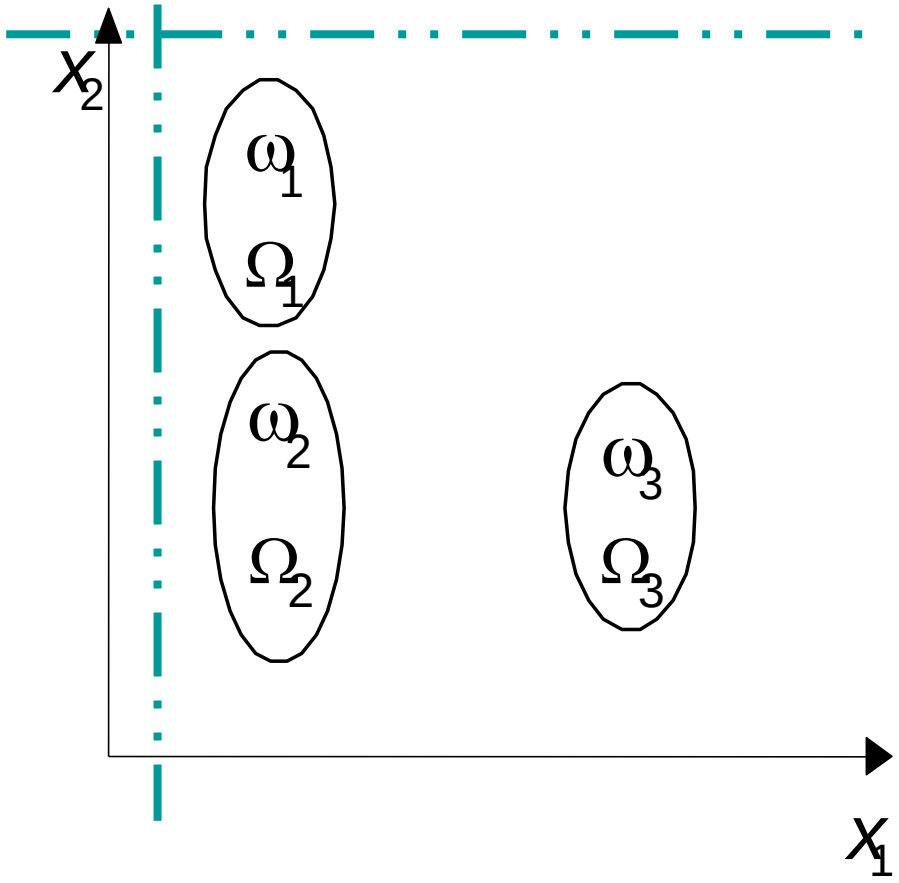
- 如果选择X_2作为分类特征，则在重叠区无法区分ω_1和ω_2这两类
- 如果选择X_1作为分类特征，则可以很容易区分ω_1和ω_2这两类



特征设计对聚类分析的影响

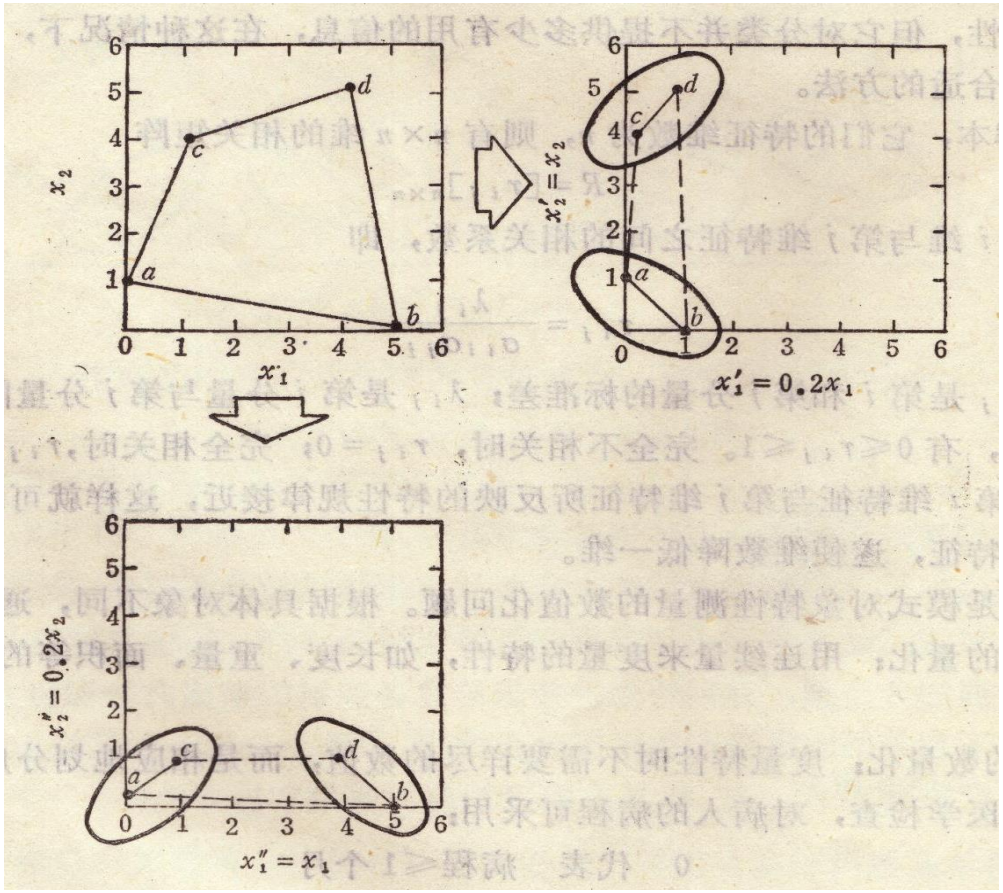
（2）特征选取不足使分类无效

- 不管选择X_1还是X_2作为分类特征，都无法单独区分这三类
- 但是，如果选择（X_1, X_2）作为分类特征，则可以区分这三类

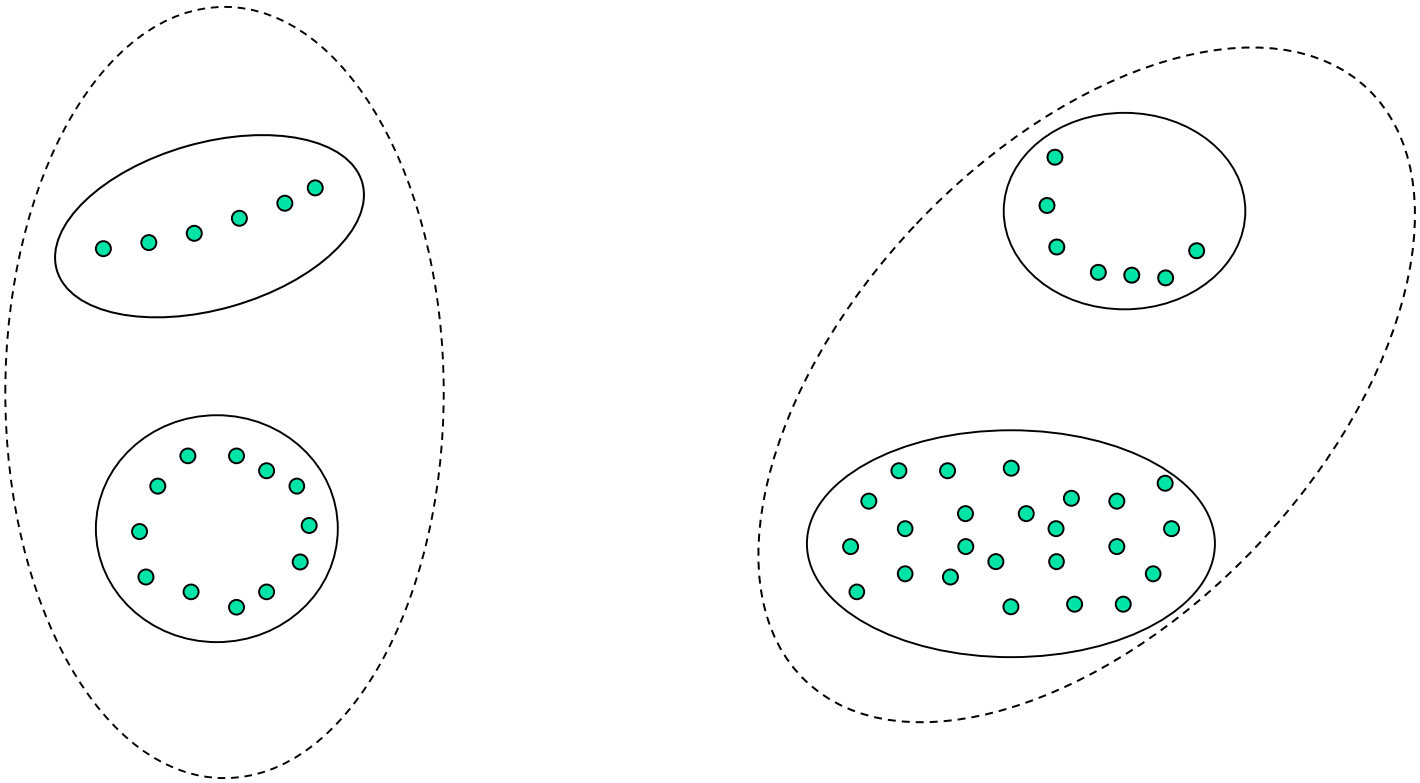


量纲不同，聚类结果可能不同

- 初始：*a,b,c,d* 属于4个不同的类（左上角）。
- 当x_1的量纲变大导致数值变小时（右上角），*a,b* 成为一类，c,*d* 成为一类。
- 当x_2的量纲变大导致数值变小时（左下角），*a,c* 成为一类，*b,d* 成为一类。
- 当然，如果x_1和x_2的量纲按照相同尺度同时变化，则不影响聚类结果



相似性（距离）测度不同，聚类结果也不同



粗聚类则为2类，细聚类则为4类

影响聚类方法有效性的几个因素

综上可见：

- 选择什么特征？
- 选择多少个特征？
- 选择什么样的量纲？
- 选择什么样的距离测度？

对分类结果都会产生极大影响。

2.2 相似性测度和聚类准则

- **相似性测度**：衡量模式之间相似性的一种量度。
- 由n个特征值组成的n维向量 $\boldsymbol{X}=[x_1,x_2,....,x_n]^{\mathrm{T}}$，称为该模式（样本）的特征矢量。它相当于特征空间中的一个点，以特征空间中的点间距离作为模式相似性的测量，作为模式归类依据，距离越小，越“相似”。
- 复习一下：

$$\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}=\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}\begin{bmatrix}y_1 & y_2 & y_3\end{bmatrix}=\begin{bmatrix}y_1^2 & y_1y_2 & y_1y_3\\y_2y_1 & y_2^2 & y_2y_3\\y_3y_1 & y_3y_2 & y_3^2\end{bmatrix}$$

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\begin{bmatrix}y_1 & y_2 & y_3\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}=y_1^2+y_2^2+y_3^2=\|\boldsymbol{Y}\|^2$$

1. 欧氏距离（Euclidean，欧几里德）

设$\boldsymbol{X}_1$、$\boldsymbol{X}_2$为两个n维模式样本

$$\boldsymbol{X}_1=[x_{11},x_{12},\ldots,x_{1n}]^{\mathrm{T}} \qquad \boldsymbol{X}_2=[x_{21},x_{22},\ldots,x_{2n}]^{\mathrm{T}}$$

欧氏距离定义为：

$$D(\boldsymbol{X}_1,\boldsymbol{X}_2)=\|\boldsymbol{X}_1-\boldsymbol{X}_2\|=\sqrt{(\boldsymbol{X}_1-\boldsymbol{X}_2)^{\mathrm{T}}(\boldsymbol{X}_1-\boldsymbol{X}_2)}$$

$$=\sqrt{(x_{11}-x_{21})^2+\cdots\cdots+(x_{1n}-x_{2n})^2}$$

距离越小，越相似。

注意：

- 各特征维上应当是相同的物理量；
- 注意同类物理量的量纲应该一样。

2. 马氏距离（Mahalanobis，马哈拉诺比斯）

平方表达式：$D^2=(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{X}-\boldsymbol{M})$

式中，X：模式向量；　M：均值向量；

C：该类模式总体的协方差矩阵。

对n维向量：　$\boldsymbol{X}=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$　　$\boldsymbol{M}=\begin{bmatrix}m_1\\ \vdots\\ m_n\end{bmatrix}$

$$\boldsymbol{C}=E\left\{(\boldsymbol{X}-\boldsymbol{M})((\boldsymbol{X}-\boldsymbol{M}))^{\mathrm{T}}\right\}$$

$$=E\{\begin{bmatrix}(x_1-m_1)\\ (x_2-m_2)\\ \vdots\\ (x_n-m_n)\end{bmatrix}\begin{bmatrix}(x_1-m_1) & (x_2-m_2) & \cdots & (x_n-m_n)\end{bmatrix}\}$$

2. 马氏距离

$$
= \begin{bmatrix}
E(x_1-m_1)(x_1-m_1) & E(x_1-m_1)(x_2-m_2) & \cdots & E(x_1-m_1)(x_n-m_n) \\
E(x_2-m_2)(x_1-m_1) & E(x_2-m_2)(x_2-m_2) & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots \\
E(x_n-m_n)(x_1-m_1) & \cdots & \cdots & E(x_n-m_n)(x_n-m_n)
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\
\sigma_{21}^2 & \ddots & \sigma_{jk}^2 & \vdots \\
\vdots & \vdots & \sigma_{kk}^2 & \vdots \\
\sigma_{n1}^2 & \cdots & \cdots & \sigma_{nn}^2
\end{bmatrix}
$$

马氏距离：在各分量特征维度计算样本模式与均值模式的距离上，剔除了该维度上模式类的方差影响——方差大，说明模式取值变化大，因此计算距离时要用方差归一化，这样得出的分量模式与均值模式的距离才具有比较意义。

优点：排除了模式样本之间的相关性影响 。

特例：当 $C = I$ 时，马氏距离退化为欧氏距离。

3. 明氏距离(Minkowski，闵可夫斯基)

n维模式向量X_i、X_j间的**明氏距离**表示为：

$$D_m(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left[\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|^m\right]^{1/m}$$

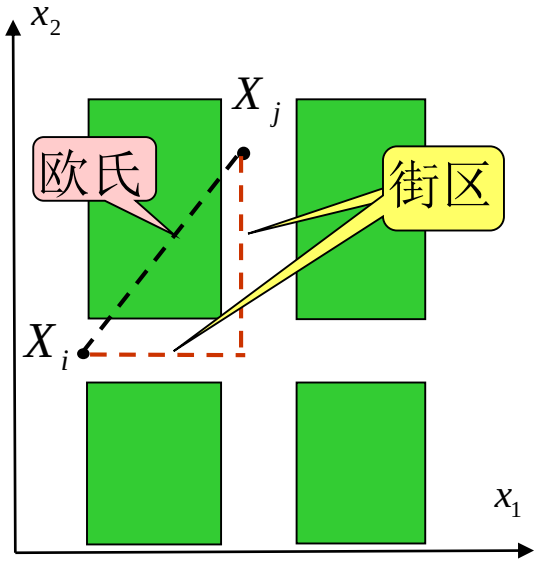
式中，x_{ik}、x_{jk}分别表示X_i和X_j的第k个分量。也称为 m-范数。

当m=2时，明氏距离为欧氏距离。

当m=1时：

$$D_1(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|$$

称为**“街区”距离** (“City block” distance)、**曼哈顿距离**。



示例：当k=2时：图中

$$D_1(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|$$

4．汉明距离（Hamming）

设X_i、X_j为n维二值模式（分量取值1或－1）样本向量，则

汉明距离： $$D_h(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{1}{2}\left(n-\sum_{k=1}^{n}x_{ik}\cdot x_{jk}\right)$$

式中，x_{ik}、x_{jk}分别表示$\boldsymbol{X}_i$和$\boldsymbol{X}_j$的第k个分量。

说明：汉明距离中的**求和式**，表达的是两个二值向量之间，**同值分量数与不同值分量数之差**；显然，这个求和式表示的差值越大，表示两个向量越相似，于是求和式的负值也就越小，因此求和式的负值可表示距离，取值于（*-n，n*）。为更直观表达，用最大值*n*加上这个差值并除以2，取值于（*0，n*）。

———0：表示两模式向量的各个分量都相同；

———*n*：表达两模式向量的各个分量都不同；

——*n/4*：表示两模式向量中，取值相同的分量数比取值不同的分量数多*n/2*。因为此时：取值相同的分量数为*(3/4)n*，取值相同的分量数为*(1/4)n*。

5. 角度相似性函数

$$S(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j}{\|\boldsymbol{X}_i\| \cdot \|\boldsymbol{X}_j\|}$$

定义为：模式向量$\boldsymbol{X}_i$，$\boldsymbol{X}_j$之间夹角的余弦。

测度基础：以两矢量的方向是否相近作为考虑的基础，矢量长度并不重要。设

$$\vec{x} = (x_1, x_2, \ldots x_n)', \vec{y} = (y_1, y_2, \ldots y_n)'$$

注意：该测度对坐标系的旋转和尺度的缩放是不变的，但对一般的线形变换和坐标系的平移不具有不变性。

6. Tanimoto测度（广义Jaccard系数）

用于0-1型二值特征情况，定义：

$$S(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j}{\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{X}_j - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j}$$

$$= \frac{\boldsymbol{X}_i, \boldsymbol{X}_j \text{之间均具有某特征的数目}}{\boldsymbol{X}_i \text{和} \boldsymbol{X}_j \text{中有某种特征的分量总数}}$$

如果把记1为“有某种特征”，记0为“无某种特征”，则Tanimoto测度的分子表达的是两特征向量之间“都具有某种特征”的分量数，分母表达的则是，在两个向量中“至少有一个向量的分量有某种特征”的分量数。

相似性测度函数的共同点都涉及到把两个相比较的向量$\boldsymbol{X}_i$，$\boldsymbol{X}_j$的分量值组合起来，但怎样组合并无普遍有效的方法，对具体的模式分类，需视情况作适当选择。

聚类准则

聚类准则：根据相似性测度确定的、衡量模式聚类结果中得到的聚类是否满足某种优化目标的一个判断标准或方法。

确定聚类准则的两种方式：

1. 阈值准则：根据规定的距离阈值进行判断。

2. 函数准则：利用聚类准则函数进行判断。

聚类准则函数：在聚类分析中，表示聚类过程中，所产生的中间分类结果质量的一种度量函数。

聚类准则函数应是模式样本集$\{X\}$和模式类别$\{S_j, \mathrm{j}=1,2,\cdots\cdots,c\}$的函数。可使聚类分析转化为寻找准则函数极值的优化问题。一种常用的指标是误差（距离）平方和。

聚类准则

聚类准则函数：$$J=\sum_{j=1}^{c}\sum_{X\in S_j}\left\|\boldsymbol{X}-\boldsymbol{M}_j\right\|^2$$

式中：c为聚类类别的数目，

$\boldsymbol{M}_j=\frac{1}{N_j}\sum_{X\in S_j}\boldsymbol{X}$为属于 S_j 集的样本的**均值向量**，

N_j为模式类S_j 中样本数目。

J代表了分属于c个聚类类别的全部模式样本与其相应类别模式均值样本之间的误差（距离）平方和。

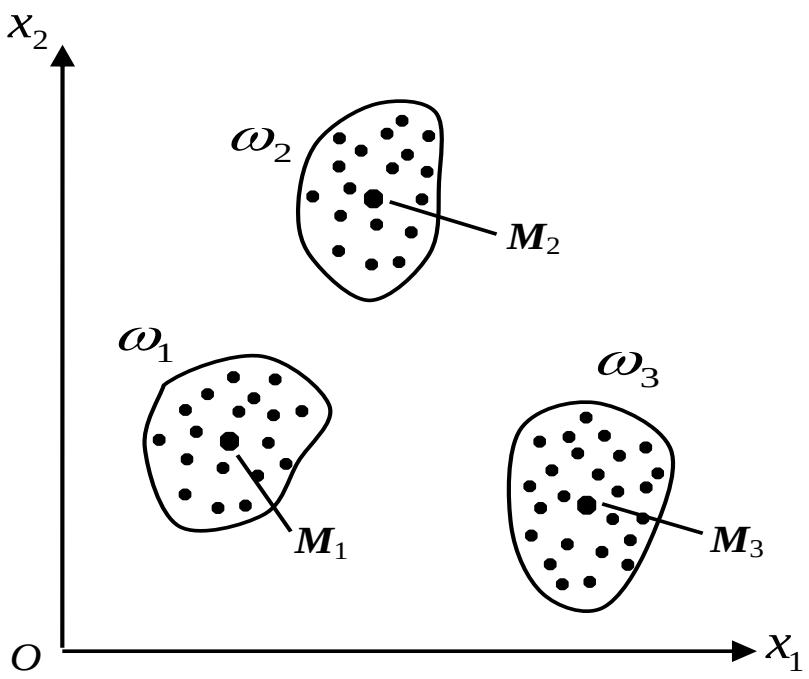
显然，这里的J是类别数c的单调减函数，因此这样的准则函数，如果不加控制，容易把任何模式集分为更多的类 。

适用范围：

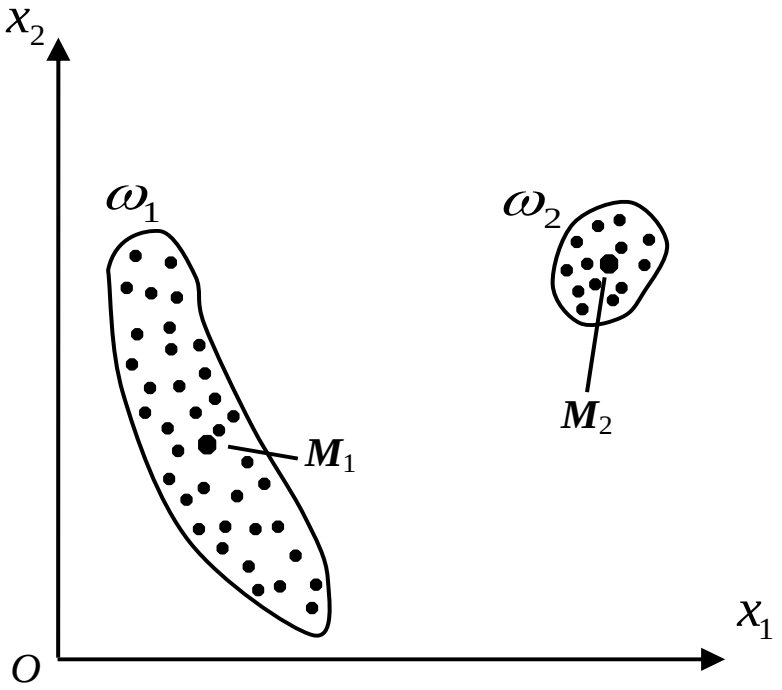
适用于各类样本密集且数目相差不多，而不同类间的样本又明显分开的情况。

聚类准则

例1：



（a）



（b）

类内误差平方和很小，类间距离很远——可得到最好的结果。

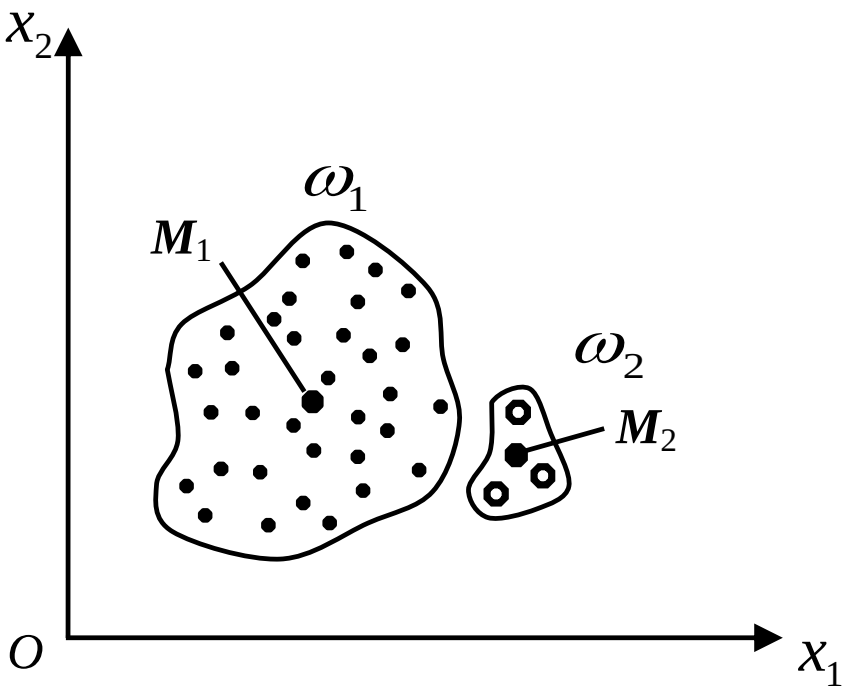
ω_1 类长轴两端距离中心很远，J 值较大，结果不易令人满意。

聚类准则

例2：另一种情况

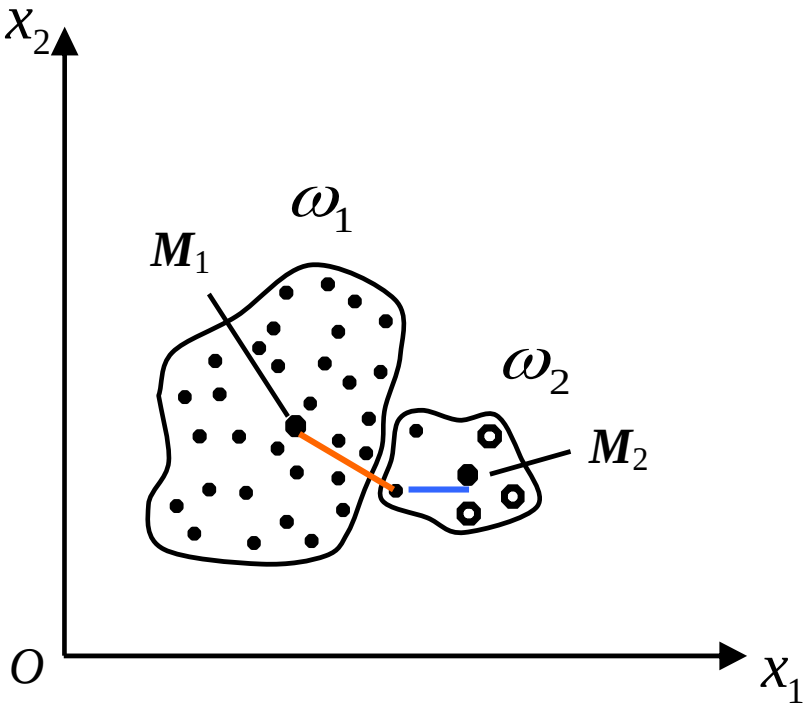
有时可能会把样本数目多的一类分拆为二，造成错误聚类。

原因：因为分拆后的J值会更小，因此对J值优化就倾向于分拆。



（a）

正确分类



（b）

错误分类

2.3 基于距离阈值的聚类算法

2.3.1 近邻聚类法

1. 问题：有N个待分类的模式$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，要求按距离阈值$T$分类到以 $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots$ 为聚类中心的模式类中。

2. 算法描述

① 任取样本$\boldsymbol{X}_i$作为第一个聚类中心的初始值，如令$\boldsymbol{Z}_1 = \boldsymbol{X}_1$。

② 计算样本$\boldsymbol{X}_2$到$\boldsymbol{Z}_1$的欧氏距离 $D_{21} = \|\boldsymbol{X}_2 - \boldsymbol{Z}_1\|$，

若 $D_{21} > T$，定义一新的聚类中心 $\boldsymbol{Z}_2 = \boldsymbol{X}_2$；

否则 $\boldsymbol{X}_2 \in$ 以$\boldsymbol{Z}_1$为中心的聚类。

2.3.1 近邻聚类法

③ 假设已有聚类中心$\boldsymbol{Z}_1$、$\boldsymbol{Z}_2$，计算 $D_{31}=\|\boldsymbol{X}_3-\boldsymbol{Z}_1\|$和 $D_{32}=\|\boldsymbol{X}_3-\boldsymbol{Z}_2\|$

若 $D_{31}>T$ 且 $D_{32}>T$ ，则建立第三个聚类中心$\boldsymbol{Z}_3=\boldsymbol{X}_3$；

否则$\boldsymbol{X}_3$∈离$\boldsymbol{Z}_1$和$\boldsymbol{Z}_2$中最近者（最近邻的聚类中心）。

……依此类推，直到将所有的N个样本都进行分类。

3. 算法特点

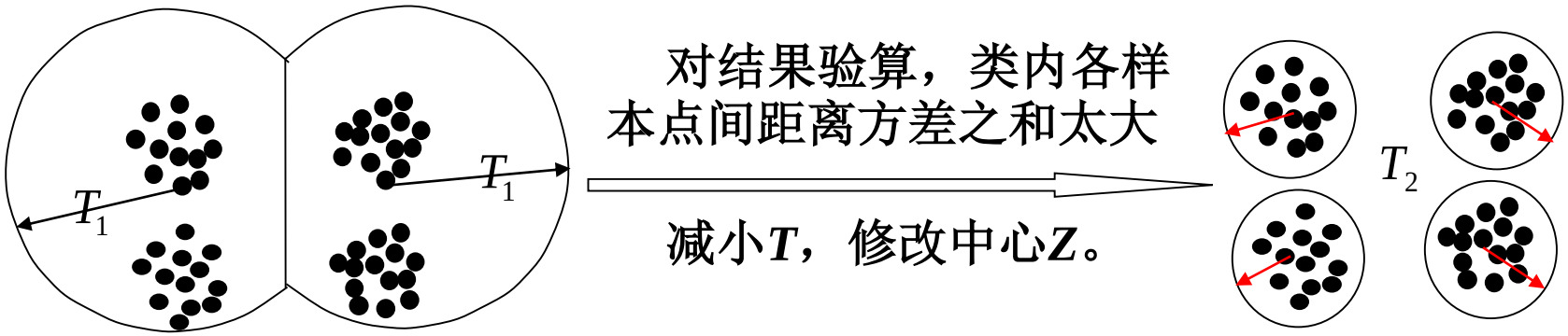
1）局限性：很大程度上依赖于第一个聚类中心的位置选择、待分类模式样本的选择次序、距离阈值T的大小以及样本分布的几何性质等。

2）优点：计算简单。（一种虽粗糙但快速的方法）

2.3.1 近邻聚类法

4. 算法讨论

用先验知识指导阈值T和起始点Z_1的选择，可获得合理的聚类结果。否则只能选择不同的初值重复试探，并对聚类结果进行验算，根据一定的评价标准，得出合理的聚类结果。



2.3.2 最大最小距离算法（小中取大距离算法）

1. 问题： 已知N个待分类的模式$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$，
分类到聚类中心$\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots$对应的类别中 。

2. 算法描述

① 选任意一模式样本做为第一聚类中心Z_1。

② 选择离Z_1距离最远的样本作为第二聚类中心Z_2。

③ 逐个计算各模式样本X_i与已确定的所有聚类中心Z_i之间的距离，并选出其中的最小距离。例如：当目前聚类中心数k=2时，计算

$$D_{i1} = \|\boldsymbol{X}_i - \boldsymbol{Z}_1\| \qquad D_{i2} = \|\boldsymbol{X}_i - \boldsymbol{Z}_2\|$$

$\min(D_{i1}, D_{i2})$，i=1，…，N　　（N个最小距离）

2.3.2 最大最小距离算法（小中取大距离算法）

④ 在所有最小距离中选出最大距离，如该最大值达到 $\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|$ 的一定分数比值（θ）以上，则相应的样本点取为新的聚类中心，返回③；否则，寻找聚类中心的工作结束。

例 $k=2$ 时

若 $\max\{\min(D_{i1},D_{i2}),i=1,2,\ldots,N\}>\theta\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|$，$0<\theta<1$

则 Z_3 存在。（θ：用试探法取为一固定分数，如1/2。）

⑤ 重复步骤③④，直到没有新的聚类中心出现为止。

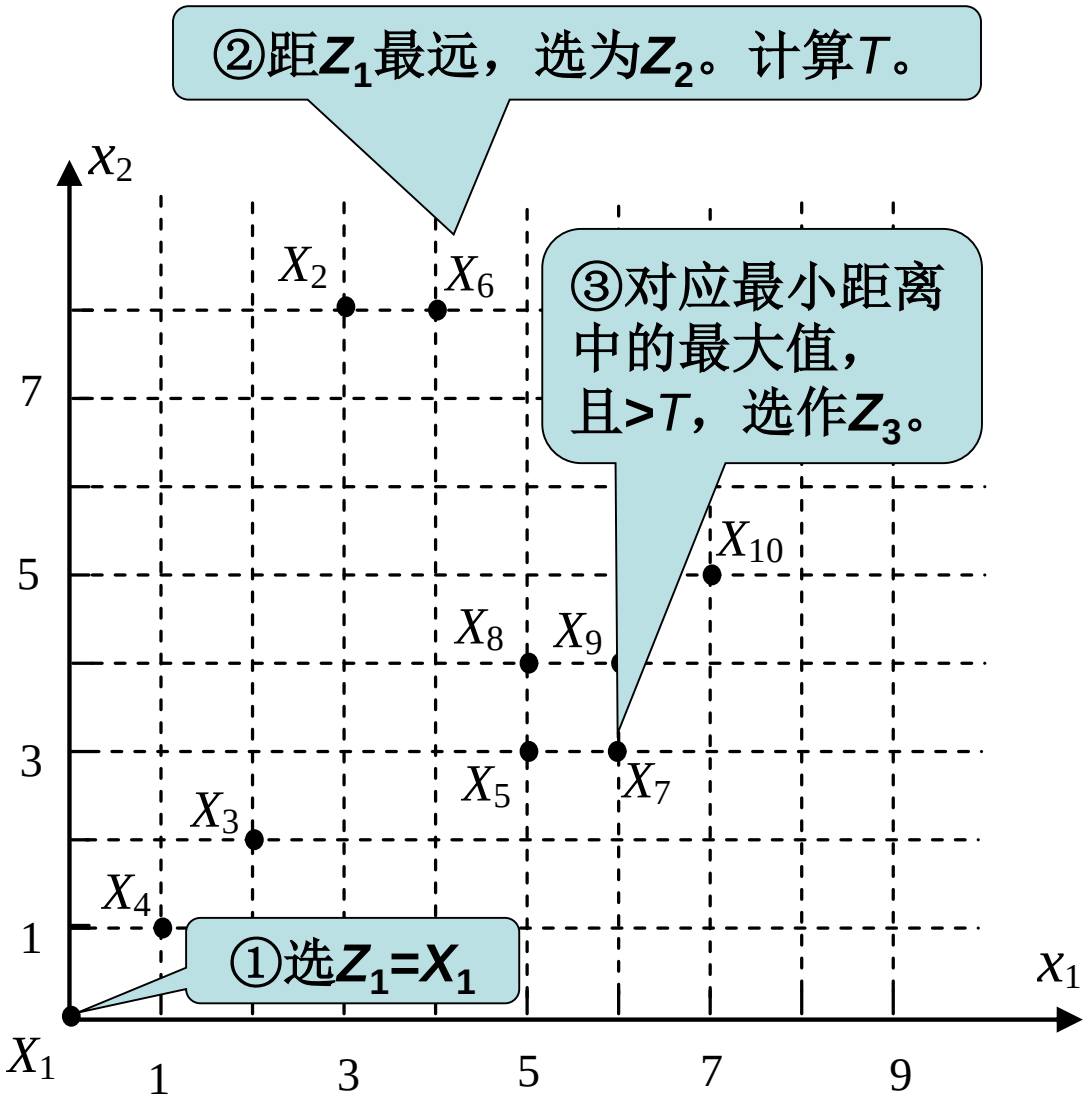
⑥ 将样本 $\{\boldsymbol{X}_i,i=1,2,\cdots,N\}$ 按最近距离划分到相应聚类中心对应的类别中。

思路总结：

二步法。先找全部中心，然后再对剩余模式归类。关键：怎样开新类，聚类中心如何定。

为使聚类中心更有代表性，**可取各类的样本均值作为聚类中心**。

例2.1 对图示模式样本用最大最小距离算法进行聚类分析



③ $T=\frac{1}{2}\|Z_1-Z_2\|=\frac{1}{2}\sqrt{80}$

10个最小距离中，X_7对应的距离>T,

$$\therefore Z_3=X_7$$

④ 用全体模式对三个聚类中心计算最小距离中的最大值，无>T 情况，停止寻找中心。

结果：$Z_1=X_1$；$Z_2=X_6$；

$Z_3=X_7$。

⑤ 对剩余模式归类（聚类）

2.4 层次聚类法（系统聚类法、谱系聚类法）

1. 算法描述

【以下讨论的是凝聚式层次聚类法，另有：分裂式层次聚类法】

1）N个初始模式样本自成一类，即建立N 类：

$$G_1(0), G_2(0), \cdots, G_N(0)$$

计算各类之间（即各样本间）的距离，得一$N \times N$维距离矩阵$\boldsymbol{D}(0)$。

“0”表示初始状态。

2.4 层次聚类法——算法描述

2）假设已求得距离矩阵$\boldsymbol{D}(n)$（n为逐次聚类合并的次数），找出$\boldsymbol{D}(n)$中的最小元素，将其对应的两类合并为一类。由此建立新的分类：

$$G_1(n+1), G_2(n+1), \cdots$$

3）再次计算：经过合并后，各个类别之间的距离，得$\boldsymbol{D}(n+1)$。
显然， $\boldsymbol{D}(n+1)$的维数是$\boldsymbol{D}(n)$的维数减一。

4）跳至第2步，重复计算及合并。

结束条件：

1）取距离阈值T，当$\boldsymbol{D}(n)$的最小分量超过给定值 T 时，算法停止。所得即为聚类结果。

2）或不设阈值T，一直将全部样本聚成一类为止，输出聚类的分级树（得到全部可能的聚类结果——谱系聚类的名称由来）。

2.4 层次聚类法——类间距离计算

2. 问题讨论：类间距离计算方法

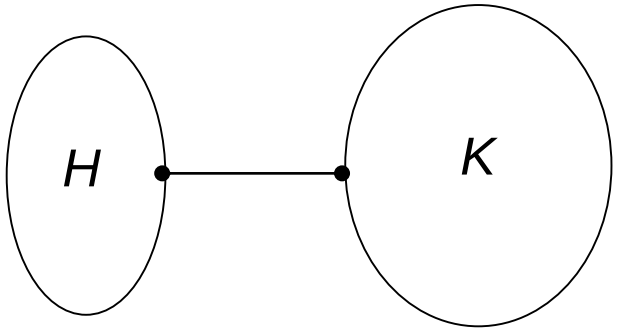
1）最短距离法

如H、K是两个聚类，则两类间的最短距离定义为：

$$D_{HK}=\min\left\{D\left(\boldsymbol{X}_H,\boldsymbol{X}_K\right)\right\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_K\in K$$

$D\left(\boldsymbol{X}_H,\boldsymbol{X}_K\right)$：$H$类中的某个样本$\boldsymbol{X}_H$和$K$类中的某个样本$\boldsymbol{X}_K$之间的欧氏距离。

D_{HK}：H类中所有样本与K类中所有样本之间的最小距离。

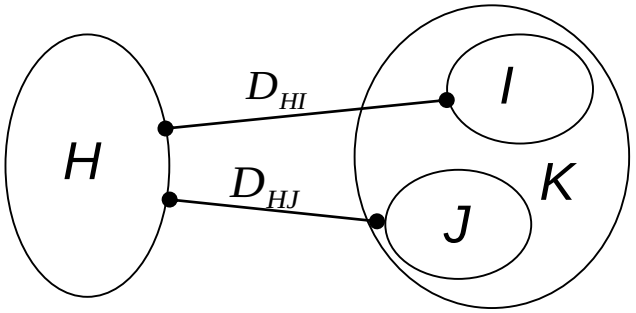


2.4 层次聚类法——类间距离计算

如果K类由I和J两类合并而成，则

$$D_{HI}=\min\{D(\boldsymbol{X}_H,\boldsymbol{X}_I)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_I\in I$$

$$D_{HJ}=\min\{D(\boldsymbol{X}_H,\boldsymbol{X}_J)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_J\in J$$



得到递推公式：　　$D_{HK}=\min\{D_{HI},D_{HJ}\}$

2）最长距离法

$$D_{HK}=\max\{D(\boldsymbol{X}_H,\boldsymbol{X}_K)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_K\in K$$

若K类由I、J两类合并而成，则

$$D_{HI}=\max\{D(\boldsymbol{X}_H,\boldsymbol{X}_I)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_I\in I$$

$$D_{HJ}=\max\{D(\boldsymbol{X}_H,\boldsymbol{X}_J)\}\quad \boldsymbol{X}_H\in H,\boldsymbol{X}_J\in J$$

有：　　$D_{HK}=\max\{D_{HI},D_{HJ}\}$

2.4 层次聚类法——类间距离计算

3）中间距离法

介于最长与最短的距离之间。如果*K*类由*I*类和*J*类合并而成，则*H*和*K*类之间的距离为

$$D_{HK}=\sqrt{\frac{1}{2}D_{HI}^2+\frac{1}{2}D_{HJ}^2-\frac{1}{4}D_{IJ}^2}$$

4）重心法

将每类中包含的样本数考虑进去。若*I*类中有n_I个样本，*J*类中有n_J个样本，则类与类之间的距离递推式为

$$D_{HK}=\sqrt{\frac{n_I}{n_I+n_J}D_{HI}^2+\frac{n_J}{n_I+n_J}D_{HJ}^2-\frac{n_In_J}{(n_I+n_J)^2}D_{IJ}^2}$$

2.4 层次聚类法——类间距离计算

5）类平均距离法

$$D_{HK}=\sqrt{\frac{1}{n_H n_K}\sum_{\substack{i\in H\\ j\in K}} d_{ij}^2}$$

d_{ij}^2：H类任一样本$\boldsymbol{X}_i$和K类任一样本$\boldsymbol{X}_j$之间的欧氏距离平方。

若K类由I类和J类合并产生，则递推式为

$$D_{HK}=\sqrt{\frac{n_I}{n_I+n_J}D_{HI}^2+\frac{n_J}{n_I+n_J}D_{HJ}^2}$$

定义类间距离的方法不同，分类结果会不太一致。实际问题中常用几种不同的方法，比较分类结果，从而选择一个比较切合实际的分类。

2.4 层次聚类法——示例

例：给出6个五维模式样本如下，按最短距离准则进行系统聚类分类。

$$\boldsymbol{X}_1=[0,3,1,2,0]^{\mathrm{T}} \quad \boldsymbol{X}_2=[1,3,0,1,0]^{\mathrm{T}} \quad \boldsymbol{X}_3=[3,3,0,0,1]^{\mathrm{T}}$$

$$\boldsymbol{X}_4=[1,1,0,2,0]^{\mathrm{T}} \quad \boldsymbol{X}_5=[3,2,1,2,1]^{\mathrm{T}} \quad \boldsymbol{X}_6=[4,1,1,1,0]^{\mathrm{T}}$$

解：（1）首先，将每一样本看作单独一类，得：

$$G_1(0)=\{\boldsymbol{X}_1\} \quad G_2(0)=\{\boldsymbol{X}_2\} \quad G_3(0)=\{\boldsymbol{X}_3\}$$

$$G_4(0)=\{\boldsymbol{X}_4\} \quad G_5(0)=\{\boldsymbol{X}_5\} \quad G_6(0)=\{\boldsymbol{X}_6\}$$

计算各类间欧氏距离，并填表：

$$D_{12}(0)=\|\boldsymbol{X}_1-\boldsymbol{X}_2\|=\left[(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+(x_{13}-x_{23})^2+(x_{14}-x_{24})^2+(x_{15}-x_{25})^2\right]^{1/2}$$

$$=[1+0+1+1+0]^{1/2}=\sqrt{3}$$

$$D_{13}(0)=\left[3^2+0+1+2^2+1\right]^{1/2}=\sqrt{15} \quad , \quad D_{14}(0) \quad , D_{15}(0) \quad , \quad D_{16}(0);$$

$$D_{23}(0) \ D_{24}(0) \ D_{25}(0) \ D_{26}(0) \ ; \quad D_{34}(0) \ D_{35}(0) \ D_{36}(0) \ \cdots$$

2.4 层次聚类法——示例

得距离矩阵$\boldsymbol{D}(0)$：

$\boldsymbol{D}(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_1(0)$	0					
$G_2(0)$	$*\sqrt{3}$	0				
$G_3(0)$	$\sqrt{15}$	$\sqrt{6}$	0			
$G_4(0)$	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(0)$	$\sqrt{11}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(0)$	$\sqrt{21}$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{4}$	0

（2）将最小距离$\sqrt{3}$对应的类$G_1(0)$和$G_2(0)$合并为1类，得新的分类。

$$G_{12}(1)=\{G_1(0),G_2(0)\}$$

$$G_3(1)=\{G_3(0)\} \qquad G_4(1)=\{G_4(0)\}$$

$$G_5(1)=\{G_5(0)\} \qquad G_6(1)=\{G_6(0)\}$$

计算聚类后的距离矩阵$\boldsymbol{D}(1)$：

由$\boldsymbol{D}(0)$ 递推出$\boldsymbol{D}(1)$ 。

2.4 层次聚类法——示例

$D(0)$	$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_5(0)$	$G_6(0)$
$G_1(0)$	0					
$G_2(0)$	$*\sqrt{3}$	0				
$G_3(0)$	$\sqrt{15}$	$\sqrt{6}$	0			
$G_4(0)$	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(0)$	$\sqrt{11}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(0)$	$\sqrt{21}$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{4}$	0

$D(1)$	$G_{12}(1)$	$G_3(1)$	$G_4(1)$	$G_5(1)$	$G_6(1)$
$G_{12}(1)$	0				
$G_3(1)$	$\sqrt{6}$	0			
$G_4(1)$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_5(1)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_6(1)$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$*\sqrt{4}$	0

（3）将$D(1)$中最小值$\sqrt{4}$对应的类合为一类，得$D(2)$。

$D(2)$	$G_{12}(2)$	$G_3(2)$	$G_4(2)$	$G_{56}(2)$
$G_{12}(2)$	0			
$G_3(2)$	$\sqrt{6}$	0		
$G_4(2)$	$\sqrt{5}$	$\sqrt{13}$	0	
$G_{56}(2)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0

2.4 层次聚类法——示例

（4）将**D**(2)中最小值$\sqrt{5}$对应的类合为一类，得**D**(3)。

D(2)	$G_{12}(2)$	$G_3(2)$	$G_4(2)$	$G_{56}(2)$
$G_{12}(2)$	0			
$G_3(2)$	$\sqrt{6}$	0	$\sqrt{13}$	
$G_4(2)$	*$\sqrt{5}$	$\sqrt{13}$	0	
$G_{56}(2)$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0

D(3)	$G_{124}(3)$	$G_3(3)$	$G_{56}(3)$
$G_{124}(3)$	0		
$G_3(3)$	$\sqrt{6}$	0	
$G_{56}(3)$	$\sqrt{7}$	$\sqrt{6}$	0

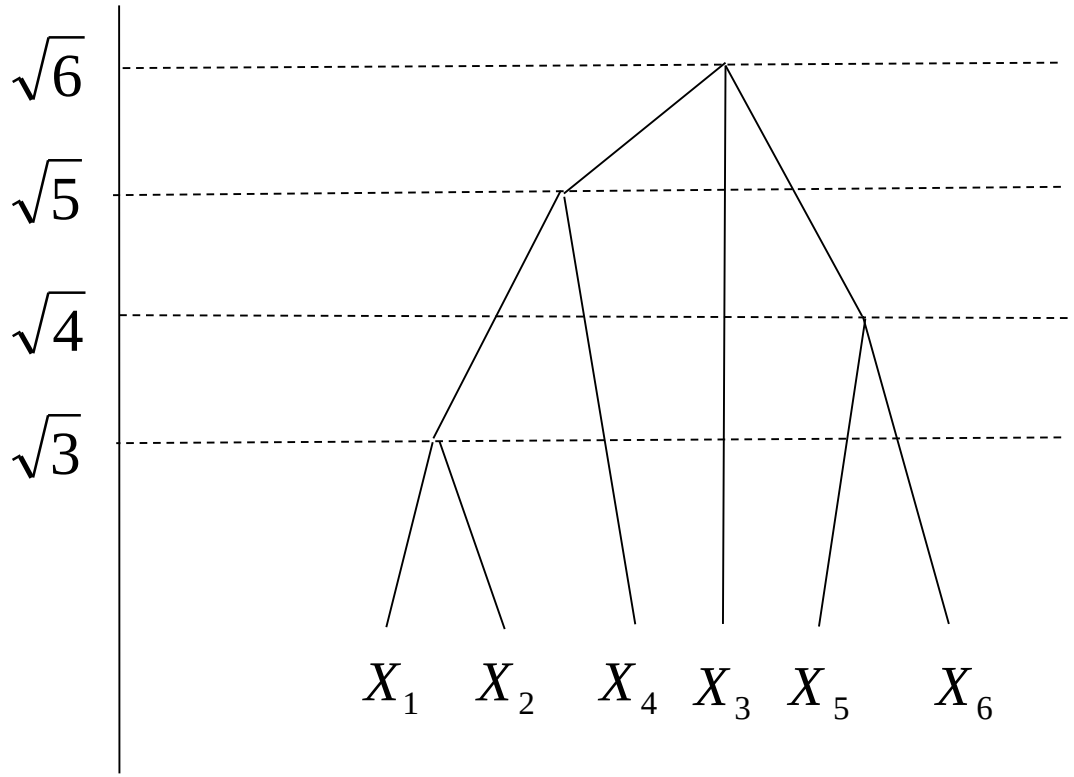
若给定阈值为$T=\sqrt{5}$，**D**(3)中的最小元素$\sqrt{6}>T$，聚类结束。

$$G_1=\{X_1,X_2,X_4\} \qquad G_2=\{X_3\} \qquad G_3=\{X_5,X_6\}$$

若无阈值，继续分下去，最终全部样本归为一类。

可给出聚类过程的树状表示图。

2.4 层次聚类法——示例



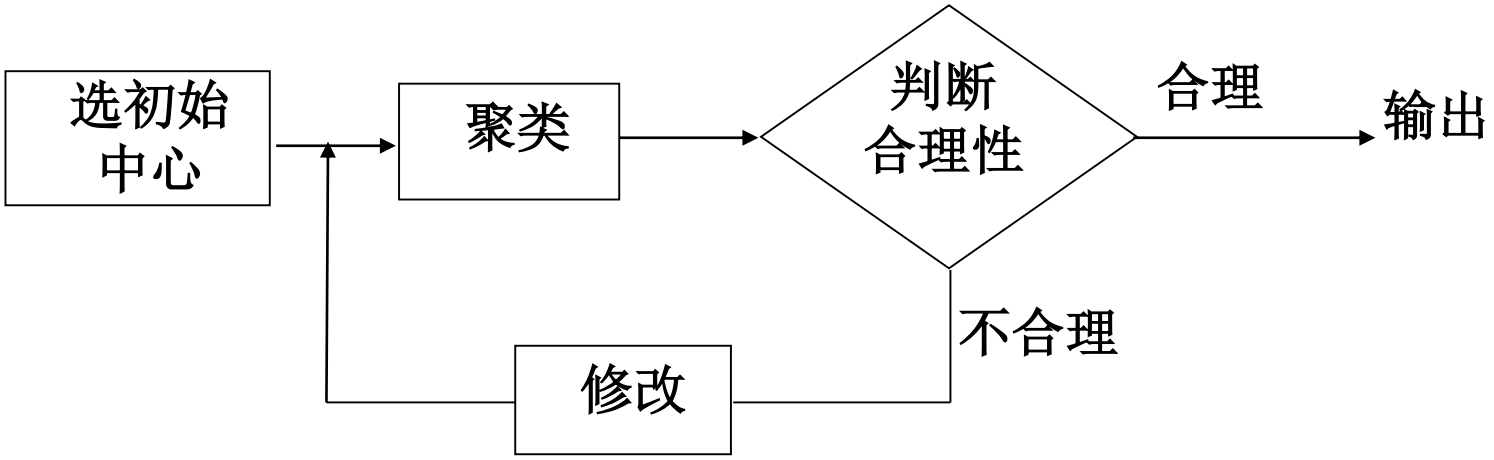
层次聚类法的树状表示

类间距离

阈值增大，

分类变粗。

2.5 动态聚类法



两种常用算法：

* ***K*-均值算法**（或C-均值算法）
* **ISODATA算法**（迭代自组织数据分析算法）

（Iterative Self-Organizing Data Analysis Techniques Algorithm）

2.5.1 *K*-均值算法

基于使聚类准则函数最小化，其准则函数：

聚类集中各点到该**聚类集中心**的距离平方和。

对于第j个聚类集，准则函数定义为

$$J_j = \sum_{i=1}^{N_j} \| \boldsymbol{X}_i - \boldsymbol{Z}_j \|^2, \quad \boldsymbol{X}_i \in S_j$$

S_j：第j个聚类集（域），聚类中心为$\boldsymbol{Z}_j$；

N_j：第j个聚类集S_j中所包含的样本个数。

对所有K个模式类有

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \| \boldsymbol{X}_i - \boldsymbol{Z}_j \|^2, \quad \boldsymbol{X}_i \in S_j$$

K-均值算法的聚类准则：聚类中心的选择，应使准则函数J极小，

也即：使J_j的值极小。

2.5.1 *K*-均值算法

应有 $$\frac{\partial J_j}{\partial \boldsymbol{Z}_j}=0$$

即 $$\frac{\partial}{\partial \boldsymbol{Z}_j}\sum_{i=1}^{N_j}\| \boldsymbol{X}_i-\boldsymbol{Z}_j \|^2=\frac{\partial}{\partial Z_j}\sum_{i=1}^{N_j}(\boldsymbol{X}_i-\boldsymbol{Z}_j)^{\mathrm{T}}(\boldsymbol{X}_i-\boldsymbol{Z}_j)=0$$

可解得 $$\boldsymbol{Z}_j=\frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{X}_i,\quad \boldsymbol{X}_i\in S_j$$

上式表明，S_j类的聚类中心应选为该类样本的均值向量。

1. 算法描述

（1）任选*K*个初始聚类中心：$\boldsymbol{Z}_1(1)$，$\boldsymbol{Z}_2(1)$，…，$\boldsymbol{Z}_K(1)$

括号内序号：迭代运算的次序号。

2.5.1 *K*-均值算法

（2）按最小距离原则将其余样本分配到*K*个聚类中心中的某一个，即：

若 $\min\left\{\left\|\boldsymbol{X}-\boldsymbol{Z}_i(k)\right\|, i=1,2,\cdots,K\right\}=\left\|\boldsymbol{X}-\boldsymbol{Z}_j(k)\right\|=D_j(k)$，则 $X\in S_j(k)$

注意：*k*——迭代运算次序号；*K*——聚类中心的个数 。

（3）再次计算各个聚类中心的新向量值 $\boldsymbol{Z}_j(k+1)\quad j=1,2,\cdots,K$

$$\boldsymbol{Z}_j(k+1)=\frac{1}{N_j}\sum_{X\in S_j(k)}\boldsymbol{X}\qquad j=1,2,\cdots,K$$

N_j：第*j*类的样本数。

这里：分别计算*K*个聚类中的样本均值向量，故称*K*-均值算法。

（4）如果 $\boldsymbol{Z}_j(k+1)\neq\boldsymbol{Z}_j(k)\quad j=1,2,\cdots,K$，则回到（2），将模式样本逐个重新分类，重复迭代计算。

如果 $\boldsymbol{Z}_j(k+1)=\boldsymbol{Z}_j(k)\quad j=1,2,\cdots,K$ ，算法收敛，计算完毕。

2.5.1 K-均值算法

“动态”聚类法

聚类过程中，
聚类中心位置或个数不断发生变化。

2. 算法讨论

结果很容易受到所选**聚类中心个数**和其**初始位置**，以及**模式样本的几何性质**及**读入次序**等的**影响**。

实际应用中，需要试探不同的K值和选择不同的聚类中心起始值。

2.5.1 K-均值算法——示例

例2.3：已知20个模式样本如下，试用K-均值算法分类。

$$\boldsymbol{X}_1=[0,0]^{\mathrm{T}}\quad \boldsymbol{X}_2=[1,0]^{\mathrm{T}}\quad \boldsymbol{X}_3=[0,1]^{\mathrm{T}}\quad \boldsymbol{X}_4=[1,1]^{\mathrm{T}}$$
$$\boldsymbol{X}_5=[2,1]^{\mathrm{T}}\quad \boldsymbol{X}_6=[1,2]^{\mathrm{T}}\quad \boldsymbol{X}_7=[2,2]^{\mathrm{T}}\quad \boldsymbol{X}_8=[3,2]^{\mathrm{T}}$$
$$\boldsymbol{X}_9=[6,6]^{\mathrm{T}}\quad \boldsymbol{X}_{10}=[7,6]^{\mathrm{T}}\quad \boldsymbol{X}_{11}=[8,6]^{\mathrm{T}}\quad \boldsymbol{X}_{12}=[6,7]^{\mathrm{T}}$$
$$\boldsymbol{X}_{13}=[7,7]^{\mathrm{T}}\quad \boldsymbol{X}_{14}=[8,7]^{\mathrm{T}}\quad \boldsymbol{X}_{15}=[9,7]^{\mathrm{T}}\quad \boldsymbol{X}_{16}=[7,8]^{\mathrm{T}}$$
$$\boldsymbol{X}_{17}=[8,8]^{\mathrm{T}}\quad \boldsymbol{X}_{18}=[9,8]^{\mathrm{T}}\quad \boldsymbol{X}_{19}=[8,9]^{\mathrm{T}}\quad \boldsymbol{X}_{20}=[9,9]^{\mathrm{T}}$$

解：① 取K=2，并选：$\boldsymbol{Z}_1(1)=\boldsymbol{X}_1=[0,0]^{\mathrm{T}}\quad \boldsymbol{Z}_2(1)=\boldsymbol{X}_2=[1,0]^{\mathrm{T}}$

② 计算距离，聚类：

$\boldsymbol{X}_1$:
$$\left.\begin{aligned} D_1&=\|\boldsymbol{X}_1-\boldsymbol{Z}_1(1)\|=0\\ D_2&=\|\boldsymbol{X}_1-\boldsymbol{Z}_2(1)\|=\sqrt{(0-1)^2+(0-0)^2}=\sqrt{1}\end{aligned}\right\}\Rightarrow D_1<D_2\Rightarrow \boldsymbol{X}_1\in S_1(1)$$

$\boldsymbol{X}_2$:
$$\left.\begin{aligned} D_1&=\|\boldsymbol{X}_2-\boldsymbol{Z}_1(1)\|=\sqrt{1}\\ D_2&=\|\boldsymbol{X}_2-\boldsymbol{Z}_2(1)\|=0\end{aligned}\right\}\Rightarrow D_2<D_1\Rightarrow \boldsymbol{X}_2\in S_2(1)$$

2.5.1 K-均值算法——示例

$\boldsymbol{X}_3$:
$$\left.\begin{aligned} D_1 &= \| \boldsymbol{X}_3 - \boldsymbol{Z}_1(1) \| = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1} \\ D_2 &= \| \boldsymbol{X}_3 - \boldsymbol{Z}_2(1) \| = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2} \end{aligned}\right\} \Rightarrow D_1 < D_2 \Rightarrow \boldsymbol{X}_3 \in S_1(1)$$

$\boldsymbol{X}_4$:
$$\left.\begin{aligned} D_1 &= \| \boldsymbol{X}_4 - \boldsymbol{Z}_1(1) \| = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2} \\ D_2 &= \| \boldsymbol{X}_4 - \boldsymbol{Z}_2(1) \| = \sqrt{(1-1)^2 + (1-0)^2} = \sqrt{1} \end{aligned}\right\} \Rightarrow D_2 < D_1 \Rightarrow \boldsymbol{X}_4 \in S_2(1)$$

……，可得到：
$$S_1(1) = \{\boldsymbol{X}_1, \boldsymbol{X}_3\} \quad N_1 = 2$$
$$S_2(1) = \{\boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5, ..., \boldsymbol{X}_{20}\} \quad N_2 = 18$$

③ 计算新的聚类中：
$$\boldsymbol{Z}_1(2) = \frac{1}{N_1} \sum_{X \in S_1(1)} \boldsymbol{X} = \frac{1}{2}(\boldsymbol{X}_1 + \boldsymbol{X}_3) = \frac{1}{2}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$$

$$\boldsymbol{Z}_2(2) = \frac{1}{N_2} \sum_{X \in S_2(1)} \boldsymbol{X} = \frac{1}{18}(\boldsymbol{X}_2 + \boldsymbol{X}_4 + \cdots + \boldsymbol{X}_{20}) == \begin{bmatrix} 5.67 \\ 5.33 \end{bmatrix}$$

④ 判断：$\because \quad \boldsymbol{Z}_j(2) \neq \boldsymbol{Z}_j(1) \quad j = 1,2$ ，故返回第②步。

2.5.1 K-均值算法——示例

② 从新的聚类中心得：

$$\boldsymbol{X}_1：\quad \left.\begin{aligned} D_1 &= \| \boldsymbol{X}_1 - \boldsymbol{Z}_1(2) \| = \cdots \\ D_2 &= \| \boldsymbol{X}_1 - \boldsymbol{Z}_2(2) \| = \cdots \end{aligned}\right\} \quad \Rightarrow \boldsymbol{X}_1 \in S_1(2)$$

$$\boldsymbol{X}_{20}：\quad \left.\begin{aligned} D_1 &= \| \boldsymbol{X}_{20} - \boldsymbol{Z}_1(2) \| = \cdots \\ D_2 &= \| \boldsymbol{X}_{20} - \boldsymbol{Z}_2(2) \| = \cdots \end{aligned}\right\} \quad \Rightarrow \boldsymbol{X}_{20} \in S_2(2)$$

有：

$$S_1(2) = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_8\} \quad N_1 = 8$$

$$S_2(2) = \{\boldsymbol{X}_9, \boldsymbol{X}_{10}, ..., \boldsymbol{X}_{20}\} \quad N_2 = 12$$

③ 计算聚类中心：

$$\boldsymbol{Z}_1(3) = \frac{1}{N_1} \sum_{X \in S_1(2)} \boldsymbol{X} = \frac{1}{8}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_8) = \begin{bmatrix} 1.25 \\ 1.13 \end{bmatrix}$$

$$\boldsymbol{Z}_2(3) = \frac{1}{N_2} \sum_{X \in S_2(2)} \boldsymbol{X} = \frac{1}{12}(\boldsymbol{X}_9 + \boldsymbol{X}_{10} + \cdots + \boldsymbol{X}_{20}) == \begin{bmatrix} 7.67 \\ 7.33 \end{bmatrix}$$

2.5.1 K-均值算法——示例

④ $\because \boldsymbol{Z}_j(3) \neq \boldsymbol{Z}_j(2) \quad j=1,2$

返回第②步，以$\boldsymbol{Z}_1(3)$， $\boldsymbol{Z}_2(3)$为中心进行聚类。

② 以新的聚类中心分类，求得的分类结果与前一次迭代结果相同：

$$S_1(3)=S_1(2) \quad S_2(3)=S_2(2)$$

③ 计算新聚类中心向量值，聚类中心与前一次结果相同，即：

$$\boldsymbol{Z}_j(4)=\boldsymbol{Z}_j(3), \quad j=1,2$$

④ $\because \boldsymbol{Z}_j(4)=\boldsymbol{Z}_j(3)$ 故算法收敛，得聚类中心为

$$\boldsymbol{Z}_1=\begin{bmatrix}1.25\\1.13\end{bmatrix}, \quad \boldsymbol{Z}_2=\begin{bmatrix}7.67\\7.33\end{bmatrix}$$

结果图示：

2.5.1 K-均值算法——示例

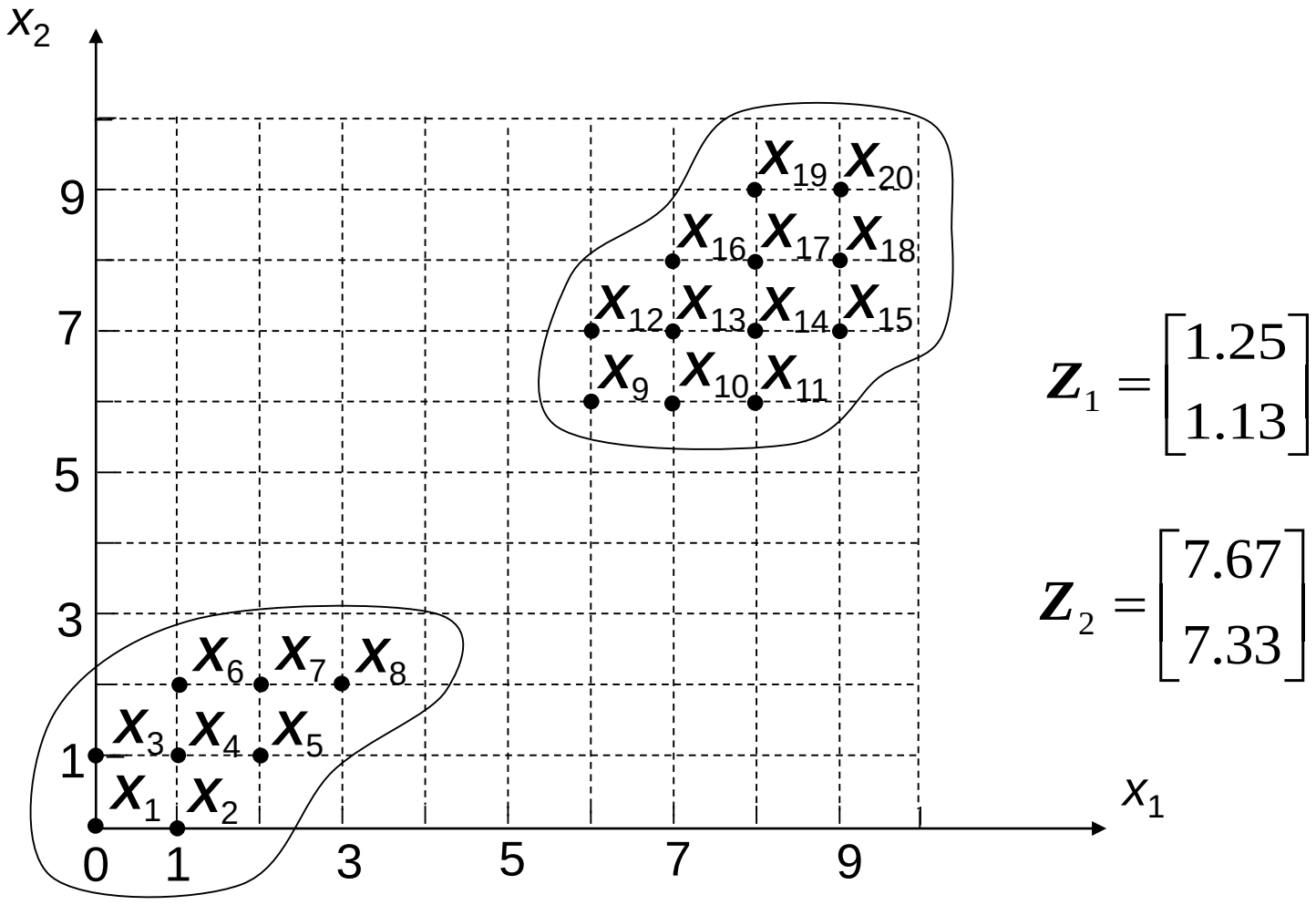
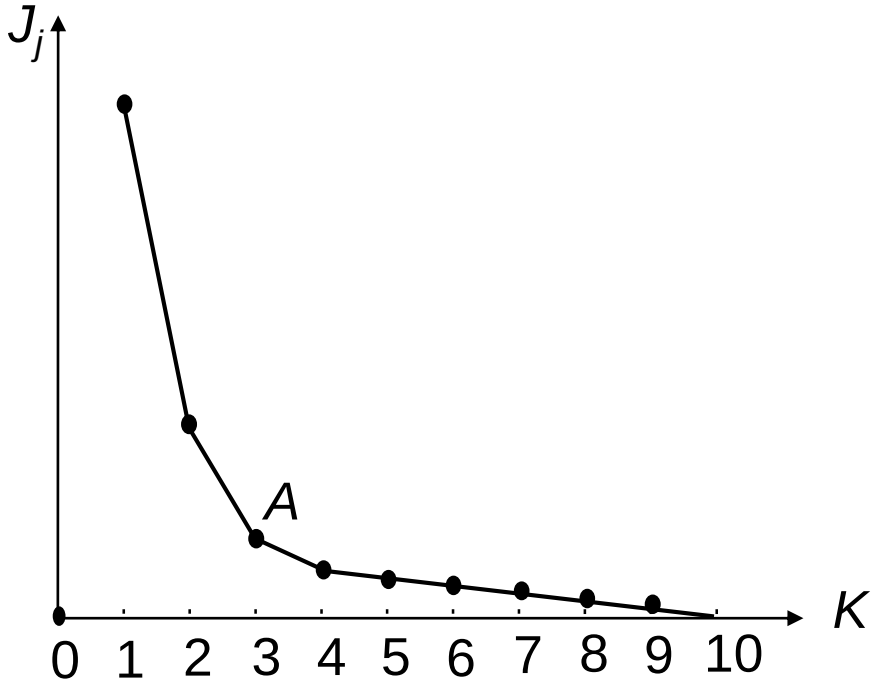


图2.10 K-均值算法聚类结果

2.5.1 K-均值算法——示例

3、聚类准则函数J_j与K的关系曲线

上述K-均值算法，其类型数目假定已知为K个。当K未知时，可以令K逐渐增加，此时J_j会单调减少。最初减小速度快，但当K增加到一定数值时，减小速度会减慢，直到K=总样本数N时，$J_j=0$。J_j—K关系曲线如下图：



曲线的拐点A对应着接近最优的K值（J值减小量、计算量以及分类效果的**权衡**）。

并非所有的情况都容易找到关系曲线的**拐点**（**肘点：导数变化率最大的点**）。迭代自组织的数据分析算法可以确定模式类的个数K。

2.6 聚类结果的评价

1、评价的重要性

1) 对高维特征向量样本，不能直观看清聚类效果。

2) 人机交互系统中，需要迅速地判断中间结果，及时指导输入参数的改变，较快地获得较好的聚类结果。

2、常用的几个指标

1) 聚类中心之间的距离——越大越好。

2) 各个聚类域中样本数目——尽可能相差不大。

3) 各个聚类域内样本的标准差向量——尽可能小。

以上都是原则性要求。实际情况下，受模式实际分布情况的影响，需要综合权衡。比如：实际情况中的样本集的确包含两个类，但是类间距离本身不大，且类内方差偏大，如果简单按聚类准则函数的极值进行优化求解，很可能把本来属于同一类的模式强行分为两类！

课后作业

- 见另文。
- 下次上课前提交。
- 最好使用电子档。

End of This Part