模式识别引论

聚类分析

赵启军

四川大学计算机学院

课程安排

- 模式识别概述（第1周）
- 贝叶斯决策理论、密度函数估计（第2-3周）
- 线性、非线性判别函数（第4-5周）
- 近邻分类器（第6周）
- 特征提取与选择（第7-8周）
- 人脸检测专题（第9周）
- 非监督学习、聚类分析（第10周）
- 统计学习理论、支持向量机（第11-12周）
- 人脸识别专题（第13周）
- 指纹识别专题（第14周）
- 习题课及课程设计专题（第15-16周）

监督vs.无监督学习

监督学习

- 训练样本的类别已知
- 利用已知类别的训练样本学习出决策面，以对不同类别的样本进行分类

无监督学习

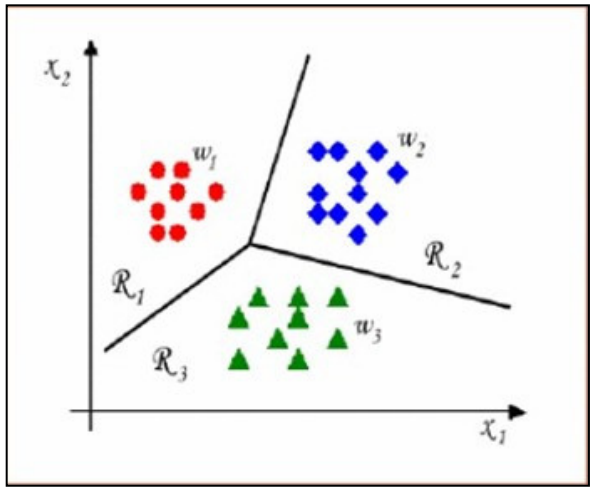
- 训练样本的类别没有标记
- 分析数据的内在规律、揭示数据的内部结构和性质，为有效地设计有针对性的分类器提供参考

主成分分析（PCA）是监督学习还是无监督学习？
我们之前学习的各种分类器设计方法呢？

分类vs.聚类

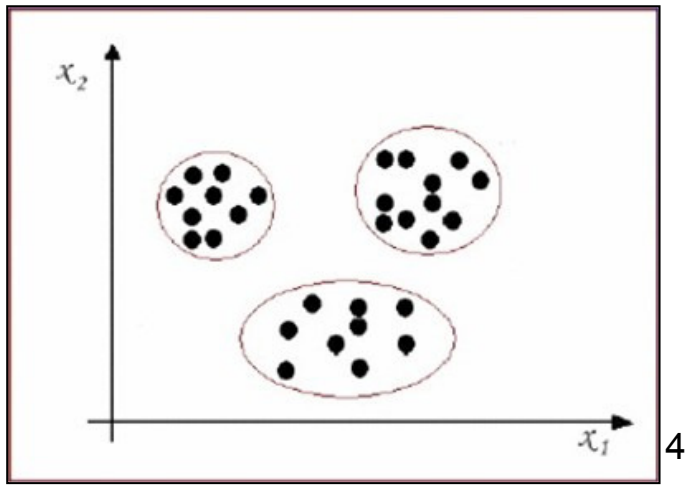
分类

- 根据已有标签的训练样本构建划分边界或对特征空间进行划分



聚类

- 挖掘没有标签的一系列样本潜在的类别

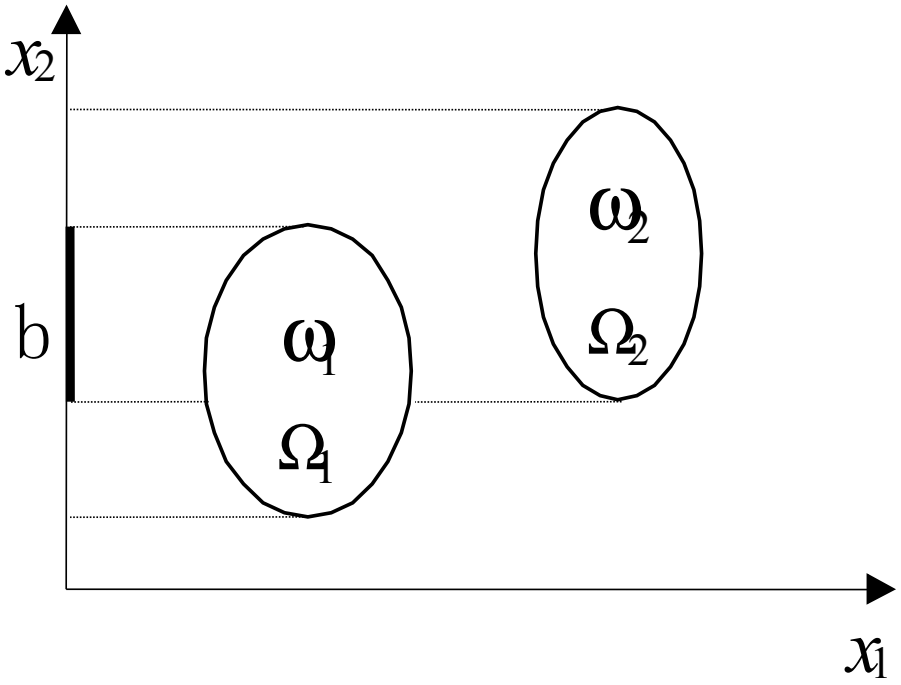


数据聚类的三个要点

- 聚类分析
  - 对一批没有标出类别的模式样本集，按照样本之间的相似程度分类，相似的归为一类，不相似的归为另一类，这种分类称为聚类分析，也称为无监督分类
- 三个要点
  - 相似性度量：如何度量样本间的相似性？
  - 聚类准则：使某种聚类准则达到极值为最佳
  - 聚类算法：用什么算法找出使准则函数取极值的最好聚类结果

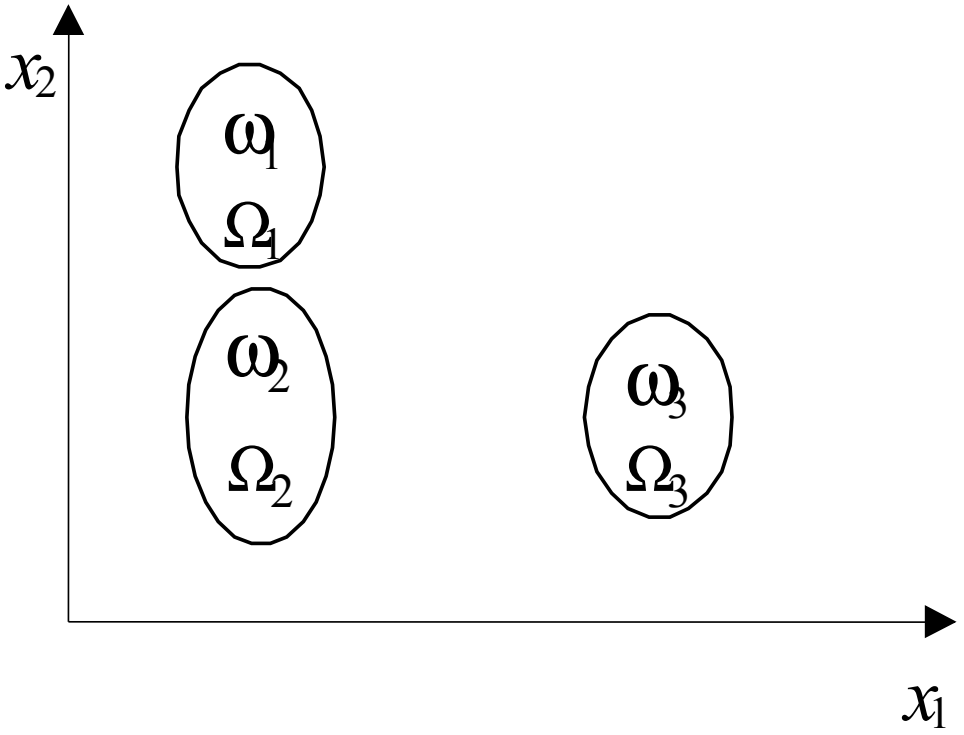
聚类的有效性

- 取决于聚类算法和特征点分布情况
- 影响聚类效果的因素
  - 特征选取不当



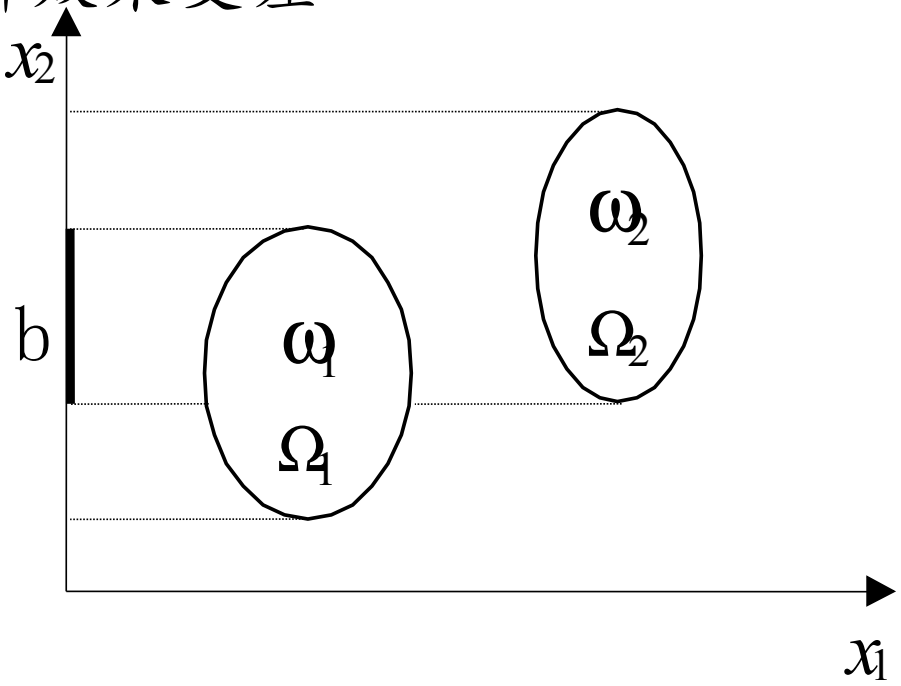
聚类的有效性

- 取决于聚类算法和特征点分布情况
- 影响聚类效果的因素
  - 特征选取不足可能使不同类别的模式判为一类



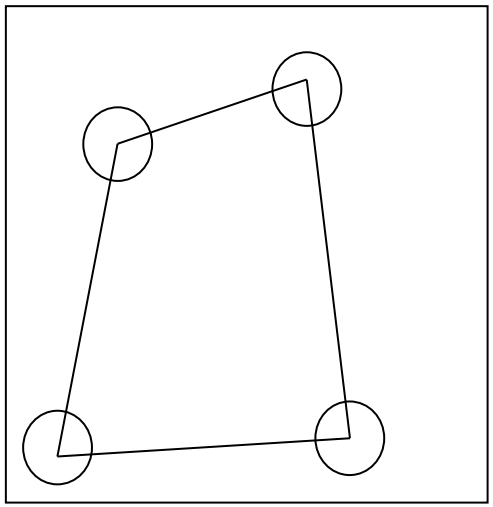
聚类的有效性

- 取决于聚类算法和特征点分布情况
- 影响聚类效果的因素
  - 特征选取过多可能无益反而有害,增加分析负担并使分析效果变差

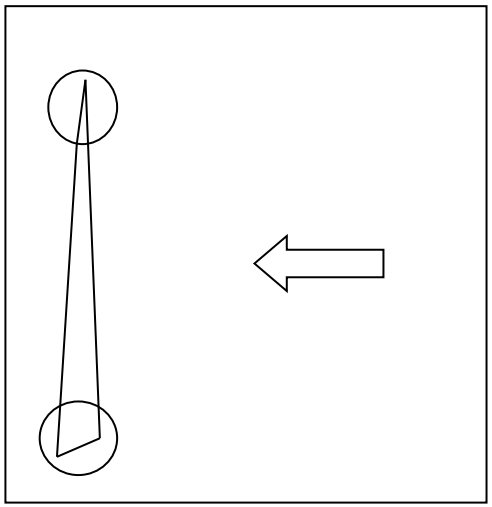


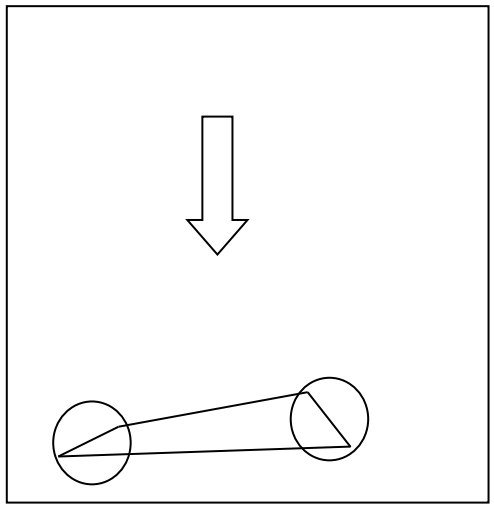
聚类的有效性

- 取决于聚类算法和特征点分布情况
- 影响聚类效果的因素
  - 量纲选取不当



两个方向的量纲一致





压缩某个方向的量纲
（如原来以厘米为单位，现改为以米为单位）

聚类的有效性

- 取决于聚类算法和特征点分布情况
- 影响聚类效果的因素
  - 选择什么特征？
  - 选择多少特征？
  - 选择什么量纲？
  - 选择什么相似性测度？

相似性测度

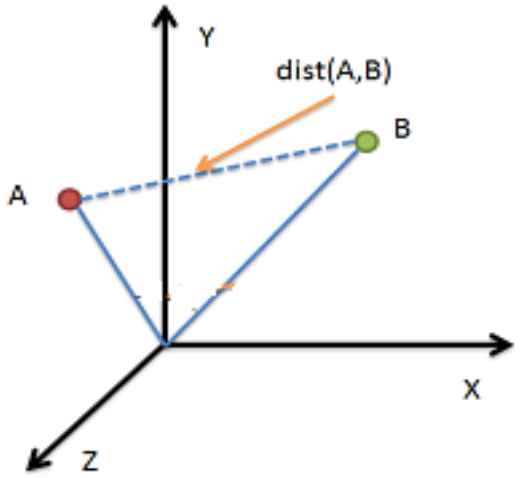
- 度量样本（或类别）之间的相似性或差异
- 聚类分析的基础
- 在计算相似度之前有时需要对数据进行规范化或者标准化（如统一量纲、去均值、规一化到指定区间等）以实现相似度计算的不变性（如不受坐标系平移的影响等）
- 常用的相似性测度
  - 距离测度
  - 相似性函数
  - 匹配测度

样本间距离测度

- 常用的样本间距离测度
  - 欧氏距离（Euclidean Distance）

$$d(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\| = [\sum_{i=1}^{n}(x_i - y_i)^2]^{1/2}$$

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



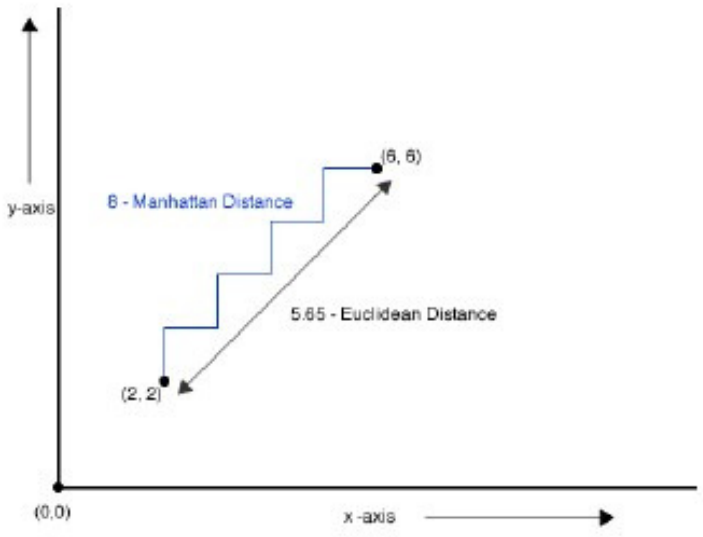
样本间距离测度

- 常用的样本间距离测度
  - 明氏距离（Minkowski Distance）

$$d(\vec{x},\vec{y})=\left[\sum_{i=1}^{n}|x_i-y_i|^m\right]^{1/m}$$

当m = 1时，又称为街坊距离或Manhattan距离

$$d(\vec{x},\vec{y})=\sum_{i=1}^{n}|x_i-y_i|$$

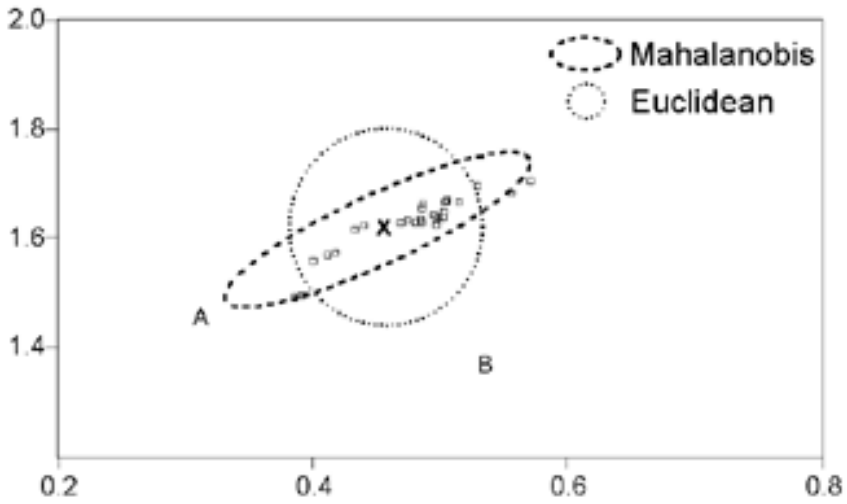


样本间距离测度

- 常用的样本间距离测度
  - 马氏距离（Mahalanobis Distance）

$$d^2(\vec{x}_i, \vec{x}_j) = (\vec{x}_i - \vec{x}_j)' V^{-1} (\vec{x}_i - \vec{x}_j)$$

其中，V 为样本协方差阵



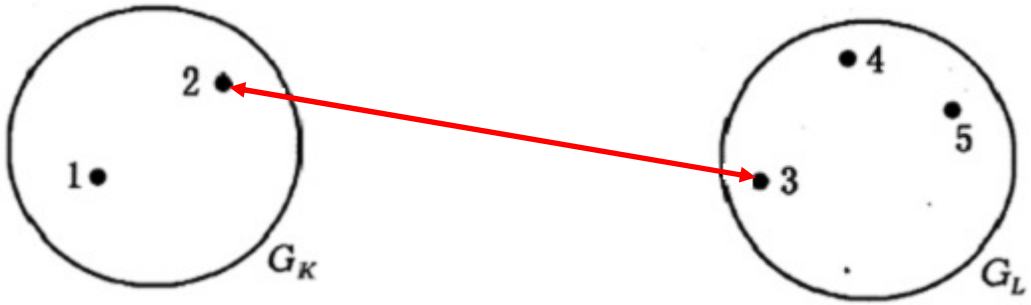
马氏距离考虑了样本特征之间的相关性，不受特征量纲选择的影响，具有平移不变性

类别间距离测度

- 常用的类别间距离测度
  - 近点距离（最近距离）：两类最近样本之间的距离

$$D_{KL} = \min\{d_{u,v} \mid u \in G_K, v \in G_L\}$$

其中，$d_{u,v}$ 表示K类中的样本u与L类中的样本v之间的距离

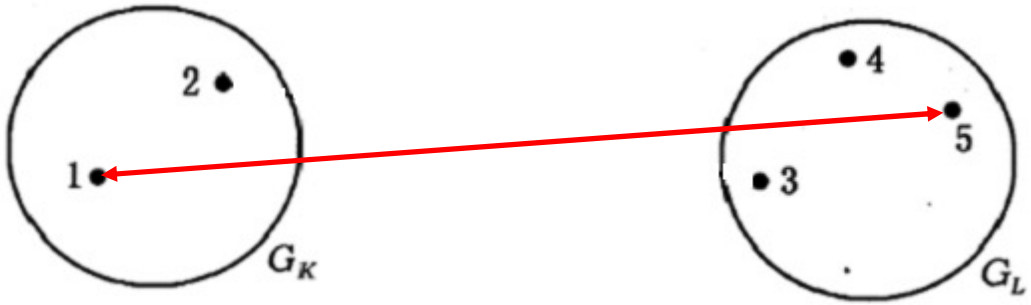


类别间距离测度

- 常用的类别间距离测度
  - 远点距离（最远距离）：两类最远样本之间的距离

$$D_{KL} = \max\left\{ d_{u,v} \mid u \in G_K, v \in G_L \right\}$$

其中，$d_{u,v}$ 表示K类中的样本u与L类中的样本v之间的距离

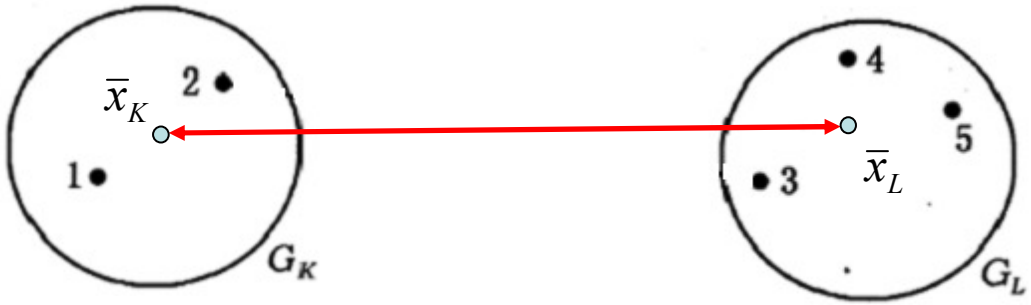


类别间距离测度

- 常用的类别间距离测度
  - 重心距离：两类均值向量之间的距离

$$D_{KL} = d(\bar{x}_K, \bar{x}_L)$$

其中，$\bar{x}_K = \frac{1}{n_K}\sum_{u \in G_K} u$ 和 $\bar{x}_L = \frac{1}{n_L}\sum_{v \in G_L} v$ 分别表示K和L类的重心（均值向量），n_K和n_L表示K类和L类的样本总数

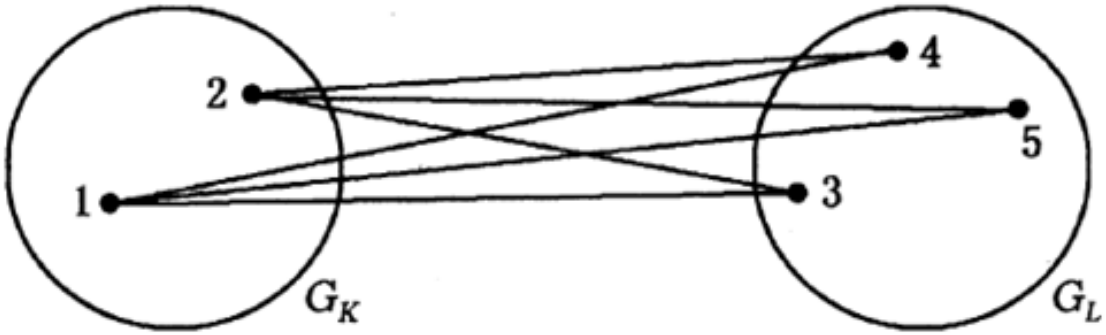


类别间距离测度

- 常用的类别间距离测度
  - 平均距离：两类所有样本之间平方距离的平均值

$$D_{KL}=\sqrt{\frac{1}{n_K n_L}\sum_{\substack{u\in G_K\\ v\in G_L}} d_{u,v}^2}$$

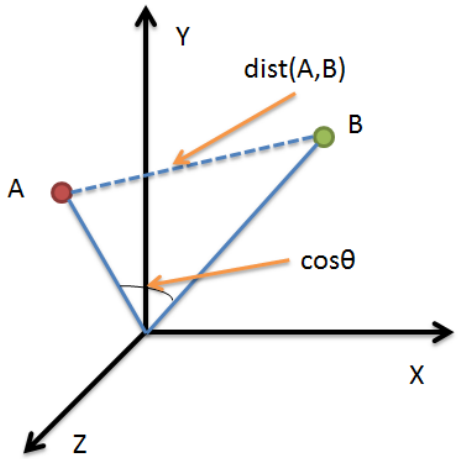
其中，$d_{u,v}$ 表示K类中的样本u与L类中的样本v之间的距离



样本间相似性函数

- 常用的样本间相似性函数
  - 角度相似函数（夹角余弦）

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}^T\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|} = \frac{\vec{x}^T\vec{y}}{\left[(\vec{x}^T\vec{x})(\vec{y}^T\vec{y})\right]^{1/2}}$$



夹角余弦对坐标系的旋转和尺度的缩放是不变的，但对坐标系的平移不具有不变性

样本间相似性函数

- 常用的样本间相似性函数
  - 相关系数：数据中心化后的夹角余弦

$$r(\vec{x},\vec{y})=\frac{(\vec{x}-\bar{\vec{x}})'(\vec{y}-\bar{\vec{y}})}{\left[(\vec{x}-\bar{\vec{x}})'(\vec{x}-\bar{\vec{x}})(\vec{y}-\bar{\vec{y}})'(\vec{y}-\bar{\vec{y}})\right]^{1/2}}$$

  - 指数相似系数

$$e(\vec{x},\vec{y})=\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\frac{3}{4}\frac{(x_i-y_i)^2}{\sigma_i^2}\right]$$

其中，σ_i 为相应特征分量的方差，n 为特征维数。
指数相似系数不受量纲影响。

样本间匹配测度

- 适用于二值特征，即特征只有两个状态（0或1），0表示无此特征、1表示有此特征
- 二值特征示例
  - 为区分猫、鸽子和金鱼，定义如下特征
    - **是否有爪子**　**是否有翅膀**　**是否生活在水里**　**是否会飞**
    - **是否有膝**　**是否有鳍**　**是否有耳廓**
  - 则这三类动物可以分别用下面的特征向量表示

    猫（1 0 0 0 1 0 1）鸽子（1 1 0 1 0 0 0）金鱼（0 0 1 0 0 1 0）

样本间匹配测度

- 对于给定的x和y中的某两个相应分量x_i与y_j

  若$x_i=1, y_j=1$ ，则称 x_i与y_j是 (1−1)匹配;

  若$x_i=1, y_j=0$ ，则称 x_i与y_j是 (1−0)匹配;

  若$x_i=0, y_j=1$ ，则称 x_i与y_j是 (0−1)匹配;

  若$x_i=0, y_j=0$ ，则称 x_i与y_j是 (0−0)匹配。

- 二值特征向量的匹配测度

  (1−1)匹配的特征数目 $a=\sum_i x_i y_i$

  (0−1)匹配的特征数目 $b=\sum_i (1-x_i) y_i$

  (1−0)匹配的特征数目 $c=\sum_i x_i (1-y_i)$

  (0−0)匹配的特征数目 $e=\sum_i (1-x_i)(1-y_i)$

样本间匹配测度

- 二值特征向量的匹配测度
  - Tanimoto测度：共同具有的特征数与各自具有的特征数之和的比值（不考虑（0-0）匹配）

$$s(\vec{x},\vec{y})=\frac{a}{a+b+c}$$

  - Rao测度：（1-1）匹配数与所选用的特征总数之比

$$s(\vec{x},\vec{y})=\frac{a}{a+b+c+e}$$

  - 简单匹配系数：（1-1）和(0-0)匹配总数与所选用特征总数之比

$$s(\vec{x},\vec{y})=\frac{a+e}{a+b+c+e}$$

聚类准则

- 如何判别某一种聚类结果的好坏？
  - 一般标准：类间距离大、类内距离小
  - 聚类的过程就是寻找某种聚类方式使得聚类准则达到最优

- 常用的聚类准则函数
  - 误差平方和准则
  - 散布准则
  - 基于模式与类核间距离的准则

误差平方和准则

- 基本思想
  - 每个类用其均值向量（又称聚类中心）代表。一个好的聚类应该使得每类中的样本到其聚类中心的距离（即误差）平方和最小。
- 误差平方和准则（也称类内距离准则）

$$\min \quad J_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left\| \vec{x}_i^{(j)} - \vec{m}_j \right\|^2$$

其中，$\vec{m}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \vec{x}_i^{(j)}$ 为第j类的聚类中心，n_j为第j类样本数目。

散布准则

- 基本思想
  - 用类内和类间散度矩阵来衡量聚类的好坏，可以同时反映同类样本的聚集程度以及不同类样本间的分离程度
- 散度矩阵

子类散度矩阵 $S_W^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}(\vec{x}_i^{(j)} - \vec{m}_j)(\vec{x}_i^{(j)} - \vec{m}_j)^T \qquad (j = 1,2,\cdots,c)$

类内散度矩阵 $S_W = \sum_{j=1}^{c}\frac{n_j}{N}S_W^{(j)}$ N为各类样本总数

类间散度矩阵 $S_B = \sum_{j=1}^{c}\frac{n_j}{N}(\vec{m}_j - \vec{m})(\vec{m}_j - \vec{m})^T \qquad \vec{m} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}_i$

总散度矩阵 $S_T = S_W + S_B$

散布准则

- 基于散度矩阵的聚类准则
  - 基本目的： $\mathrm{Tr}[S_B] \Rightarrow \max$ $\mathrm{Tr}[S_W] \Rightarrow \min$

- 利用线性代数中矩阵迹和行列式的性质，可以定义以下聚类准则

$$\max\ J_1 = \mathrm{Tr}\left[S_W^{-1} S_B\right] \qquad \max\ J_2 = \left|S_W^{-1} S_B\right|$$

$$\max\ J_3 = \mathrm{Tr}\left[S_W^{-1} S_T\right] \qquad \max\ J_4 = \left|S_W^{-1} S_T\right|$$

基于模式与类核间距离的准则

- 只用聚类中心代表类中的全部样本
  - 不能充分反映该类的模式分布结构
  - 会损失很多有用信息

- 一种解决方案：用类核代替类中心来构造聚类准则
  - 类核可以是一个函数、一个点集或其他适当的模型
  - 如马氏距离

聚类算法

- 分级聚类
  - 又称为系统聚类法或层次聚类法
  - 基本思想：按事物的相似性，或内在联系将其组织起来，组成有层次的结构，使得本质上最接近的（按最小距离准则）划为一类
- 动态聚类
  - 按准则函数确定类中心以使得准则函数取极值
  - 模式的类别和类的中心在算法运行过程中不断修正
  - 代表算法：K均值（K means）法，ISODATA法

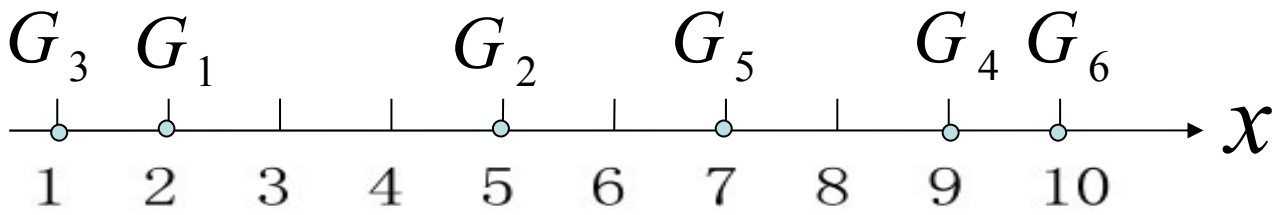
分级聚类算法

- 两种基本思路
  - 聚合法
    - 把所有样本各自看为一类，逐级聚合成一类
    - 基本思路是根据类间相似性大小逐级聚合，每级只把相似性最大的两类合并成一类，最终把所有样本聚合为一类
  - 分解法
    - 把所有样本看做一类，逐级分解为每个样本一类
    - 先把所有的对象(样本或变量)作为一大类，然后度量对象之间的距离或相似程度，并将距离或相似程度最远的对象分离出去，形成两大类(其中的一类只有一个对象)；对剩余样本不断进行分解，直至每个样本自成一类为止

聚合法分级聚类

- 假设事先指定聚类数C
- 基本步骤
  1. 设$C^*=N$，$D_i=\{x_i\}$，$i=1,2,\cdots,N$
  2. 若$C^*<=C$，则算法停止
  3. 找最近的两个类D_i和D_j
  4. 将D_i和D_j合并以代替D_i，删去D_j，C^*减1
  5. 转步骤2

聚合法分级聚类：示例



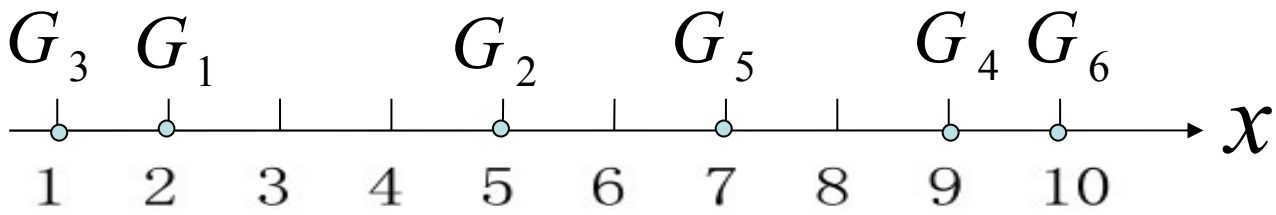
- 将一维空间中的六个样本利用聚合法进行聚类
  - 设全部样本分为6类
  - 计算距离矩阵D(0)
  - 求最小元素，并进行类合并

$\omega_7 = \{\omega_1, \omega_3\}$、$\omega_8 = \{\omega_4, \omega_6\}$

D(0)

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_2	3				
ω_3	1	4			
ω_4	7	4	8		
ω_5	5	2	6	2	
ω_6	8	5	9	1	3

聚合法分级聚类：示例



- 将一维空间中的六个样本利用聚合法进行聚类
  - 若类别数已达要求，则停止

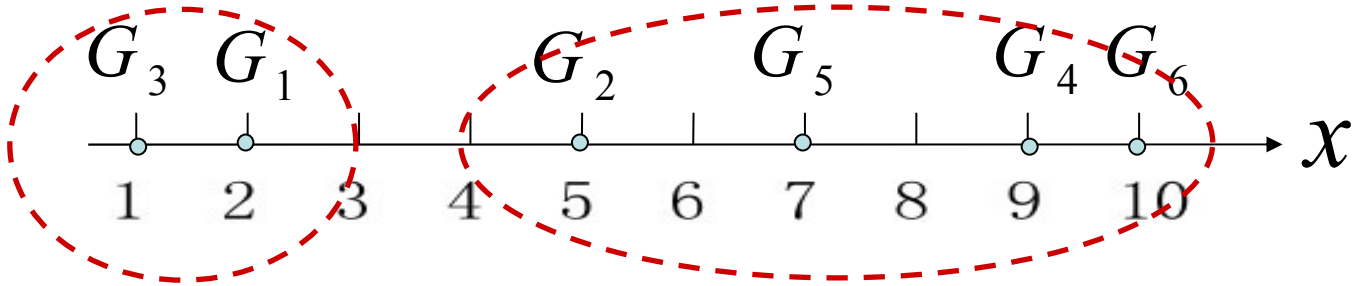
    $\omega_7=\{\omega_1, \omega_3\}$、$\omega_2$、$\omega_5$、$\omega_8=\{\omega_4, \omega_6\}$
  - 否则，继续合并最近的子类

    $\omega_9=\{\omega_2, \omega_4, \omega_5, \omega_6\}$

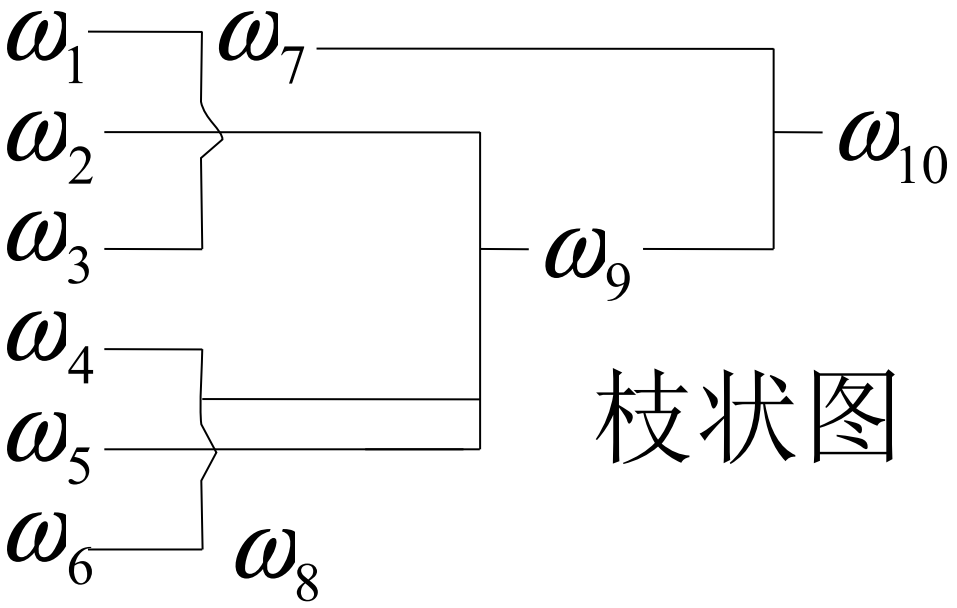
D(1)

	ω_7	ω_2	ω_8
ω_2	3		
ω_8	7	4	
ω_5	5	2	2

聚合法分级聚类：示例



- 聚合法聚类的枝状图



枝状图

动态聚类算法

- 分级聚类算法根据最小距离准则将样本进行划分，可以生成树状聚类图，方便直观地分析和选择聚类数目；但是计算量较大，对大数据聚类效率不高

- 动态聚类算法适用迭代算法不断修正样本的类别和各类的中心以使得聚类准则函数达到最优值；该类算法的三个要素
  - 选定某种距离度量作为样本间的相似性度量
  - 确定某个评价聚类结果质量的准则函数
  - 给定某个初始分类，然后用迭代算法找出使准则函数取极值的最好聚类结果

初始聚类中心的选取

- 根据经验选取
- 任选前C个样本点（如相距最远）作为初始聚类中心
- 将样本随机分成C类，计算每类的中心，并以其作为初始聚类中心
- 最大密度法
  - 求以每个特征点为球心、某一正数d_0为半径的球形域中特征点个数，这个数称为该点的密度。选取密度最大的特征点作为第一个初始类心Z_1，然后在与Z_1大于某个距离d的那些特征点中选取具有“最大”密度的特征点作为第二个初始类心Z_2，如此进行，选取C个初始聚类中心
- 从C-1类划分中产生C类划分问题的初始聚类中心

聚类中心的修正

- 批量修正法
  - 对选定的聚类中心按距离最近原则将样本划归到各聚类中心代表的类别，然后调整聚类中心
- 单步样本修正法
  - 取一样本，将其归入与其距离最近的那一类，并计算该类的样本均值，依此样本均值代替原来的聚类中心作为新的聚类中心，然后再取下一个样本，如此操作，直到所有样本都归属到相应的类别中为止

K-均值聚类

- 输入
  - 训练样本$\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$
  - 拟分类的类别数K
- 基本思想
  - 该方法取定K个类别和选取K个初始聚类中心，按最小距离原则将各模式分配到K类中的某一类，之后不断地计算类心和调整各模式的类别，最终使各模式到其判属类别中心的距离平方之和最小。

K-均值聚类：算法一

- 收敛条件为聚类中心稳定不变
- 基本步骤（t表示迭代次数）
  1. 选K个初始聚类中心，$z_1(1)$，$z_2(1)$，…，$z_K(1)$
  2. 按最小距离准则将所有样本分配给K个聚类中心中的其中一个
  3. 根据当前聚类结果更新聚类中心$z_j(t+1)$，$j=1, 2, \cdots, K$
  4. 若$z_j(t+1) \neq z_j(t)$，更新聚类中心，返回（2）
  5. 否则，算法结束，输出所得的K个聚类中心

K-均值聚类：算法二

- 收敛条件改为：通过不断调整聚类中心使得误差平方和准则函数取得极小值
- 基本步骤（t表示迭代次数）
  1. 给定允许误差 ε，令t=1
  2. 初始化聚类中心$w_i(t)$，i=1, 2, …, C
  3. 修正d_{ij}：即各样本到其所属聚类中心的距离
  4. 修正聚类中心$w_i(t+1)$
  5. 计算误差Je，如果Je< ε，则算法结束；否则t=t+1，转步骤3

K-均值聚类：示例

- 对如下20个样本用K-均值算法进行聚类，要求聚成两类（K=2）

  $x_1(0,0)$ $x_2(1,0)$ $x_3(0,1)$ $x_4(1,1)$

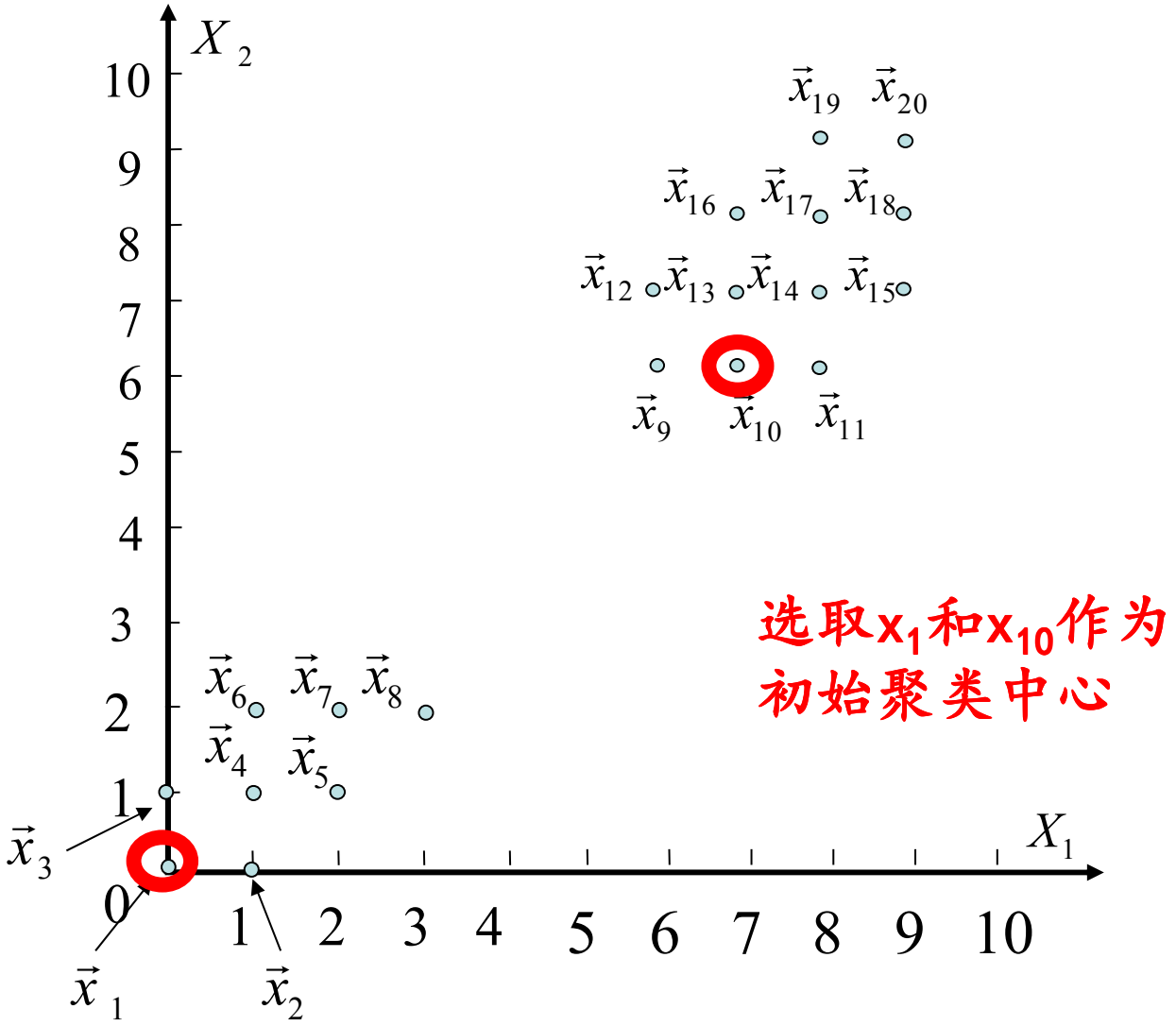
  $x_5(2,1)$ $x_6(1,2)$ $x_7(2,\ 2)$ $x_8(3,2)$

  $x_9(6,6)$ $x_{10}(7,6)$ $x_{11}(8,6)$ $x_{12}(6,7)$

  $x_{13}(7,7)$ $x_{14}(8,7)$ $x_{15}(9,7)$ $x_{16}(7,8)$

  $x_{17}(8,8)$ $x_{18}(9,8)$ $x_{19}(8,9)$ $x_{20}(9,9)$

K-均值聚类：示例


X_2
10
9
8
7
6
5
4
3
2
1
0
$\vec{x}_{19}$ $\vec{x}_{20}$
$\vec{x}_{16}$ $\vec{x}_{17}$ $\vec{x}_{18}$
$\vec{x}_{12}$ $\vec{x}_{13}$ $\vec{x}_{14}$ $\vec{x}_{15}$
$\vec{x}_9$ $\vec{x}_{10}$ $\vec{x}_{11}$
$\vec{x}_6$ $\vec{x}_7$ $\vec{x}_8$
$\vec{x}_4$ $\vec{x}_5$
$\vec{x}_3$
$\vec{x}_1$
$\vec{x}_2$
1 2 3 4 5 6 7 8 9 10
X_1
选取x_1和x_{10}作为
初始聚类中心

K-均值聚类：示例

- 计算各样本到聚类中心的距离，并按照最小距离原则进行分类

	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$z_1(1)$	1	1	2	5	5	8	13	72
$z_2(1)$	72	74	61	50	52	41	32	1

	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
$z_1(1)$	100	85	98	113	130	113	128	145	145	162
$z_2(1)$	4	2	1	2	5	4	5	8	10	13

  - 聚类结果
    - $G_1\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$
    - $G_2\{x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}\}$

K-均值聚类：示例

- 修正聚类中心

$$z_1(2) = (\frac{5}{4}, \frac{9}{8})$$

$$z_2(2) = (\frac{23}{3}, \frac{43}{6})$$

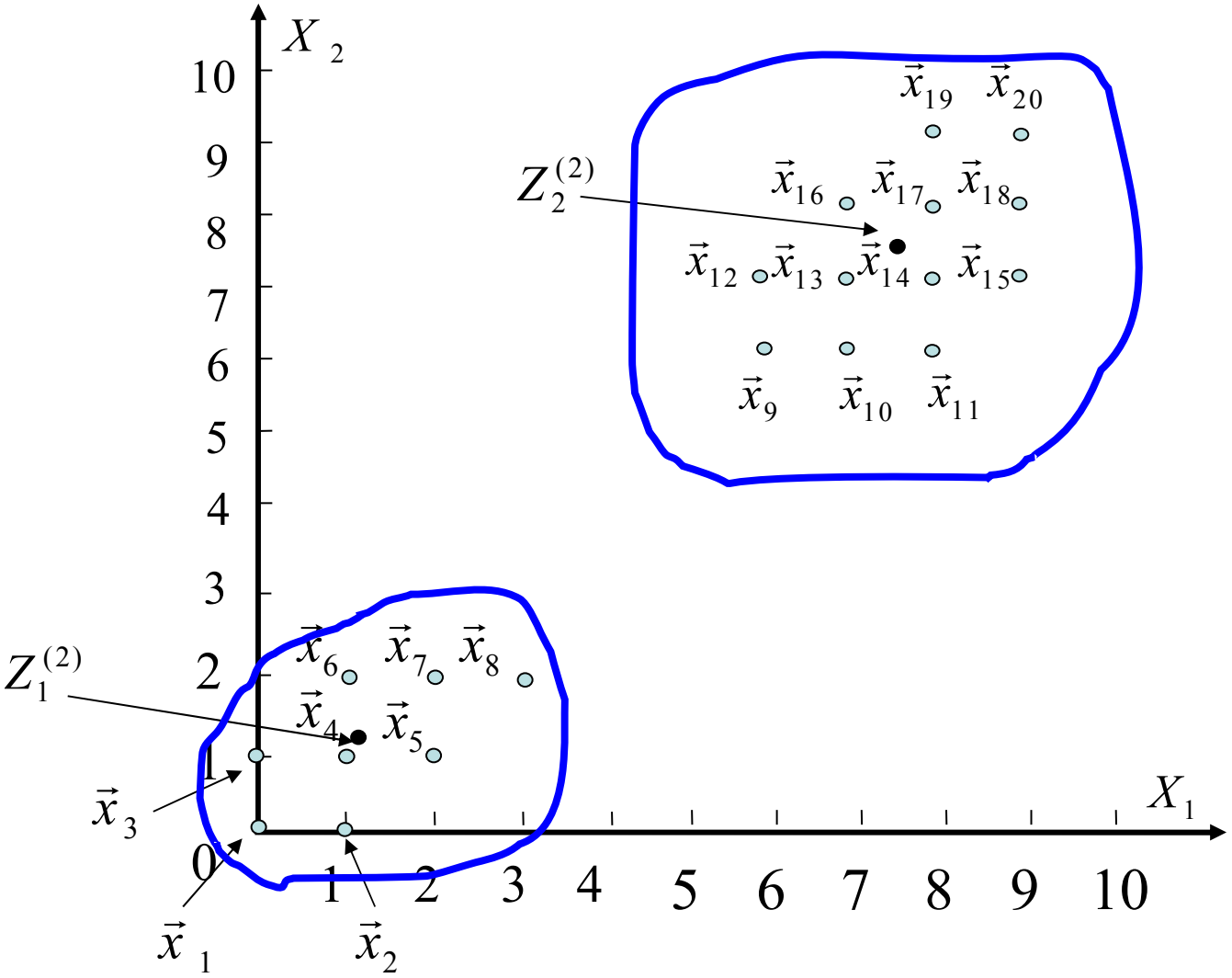
- 根据新的聚类中心重新分类样本，并再次修正聚类中心得到

$$z_1(3) = z_1(2)$$

$$z_2(3) = z_2(2)$$

- 因为聚类中心不再改变，算法结束

K-均值聚类：示例


X_2
X_1
$Z_2^{(2)}$
$Z_1^{(2)}$
$\vec{x}_{19}$ $\vec{x}_{20}$
$\vec{x}_{16}$ $\vec{x}_{17}$ $\vec{x}_{18}$
$\vec{x}_{12}$ $\vec{x}_{13}$ $\vec{x}_{14}$ $\vec{x}_{15}$
$\vec{x}_9$ $\vec{x}_{10}$ $\vec{x}_{11}$
$\vec{x}_6$ $\vec{x}_7$ $\vec{x}_8$
$\vec{x}_4$ $\vec{x}_5$
$\vec{x}_3$
$\vec{x}_1$ $\vec{x}_2$
0 1 2 3 4 5 6 7 8 9 10

K-均值聚类

- 上述示例中使用了哪一种聚类中心修正方法？
  - 批量修正法

- 如果聚类的类别数不知道呢？
  - 可以假设类别是在不断增加的，误差平方和准则函数J_e是随K的增加而减小
  - 可以通过J_e–K的关系曲线来确定合适的聚类类别数

K-均值聚类：示例二

- 中国男子足球数亚洲几流？

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	50	50	9
3	日本	28	9	4
4	韩国	17	15	3
5	伊朗	25	40	5
6	沙特	28	40	2
7	伊拉克	50	50	1
8	卡塔尔	50	40	9
9	阿联酋	50	40	9
10	乌兹别克斯坦	40	40	5
11	泰国	50	50	9
12	越南	50	50	5
13	阿曼	50	50	9
14	巴林	40	40	9
15	朝鲜	40	32	17
16	印尼	50	50	9

亚洲足球队排名

K-均值聚类：示例二

- 规格化样本数据到[0，1]区间

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	1	1	0.5
3	日本	0.3	0	0.19
4	韩国	0	0.15	0.13
5	伊朗	0.24	0.76	0.25
6	沙特	0.3	0.76	0.06
7	伊拉克	1	1	0
8	卡塔尔	1	0.76	0.5
9	阿联酋	1	0.76	0.5
10	乌兹别克斯坦	0.7	0.76	0.25
11	泰国	1	1	0.5
12	越南	1	1	0.25
13	阿曼	1	1	0.5
14	巴林	0.7	0.76	0.5
15	朝鲜	0.7	0.68	1
16	印尼	1	1	0.5

规格化后数据

K-均值聚类：示例二

- 用K-均值聚类算法将数据分成三类（K=3）
- 以日本、巴林和泰国的值作为初始聚类中心：A{0.3，0，0.19}，B{0.7，0.76，0.5}和C{1，1，0.5}
- 将各样本按最小距离原则进行分类
  - A：日本，韩国，伊朗，沙特
  - B：乌兹别克斯坦，巴林，朝鲜
  - C：中国，伊拉克，卡塔尔，阿联酋，泰国，越南，阿曼，印尼

K-均值聚类：示例二

- 按分类结果修正聚类中心
  - A：{0.21，0.4175，0.1575}
  - B：{0.7，0.7333，0.4167}
  - C：{1，0.94，0.40625}

- 根据新的聚类中心重新分类，结果没有变化，得到最终聚类结果
  - 亚洲一流：日本，韩国，伊朗，沙特
  - 亚洲二流：乌兹别克斯坦，巴林，朝鲜
  - 亚洲三流：中国，伊拉克，卡塔尔，阿联酋，泰国，越南，阿曼，印尼

ISODATA算法

- K-均值算法受初始聚类中心的选择影响大，而且类别数相对不能改变

- ISODATA（Iterative Self-Organizing Data Analysis Techniques Algorithm 迭代自组织数据分析）
  - 考虑了类别的分裂与合并，因此有了自我调整类别数的能力
  - 合并发生在某一类样本个数太少，或者两类聚类中心之间距离太小的情况
  - 分裂发生在某一类别的某分量出现类内方差过大的现象

ISODATA算法

- ISODATA聚类分析控制参数
  - C：预期的类数
  - N_c：初始聚类中心个数(可以不等于c)
  - θ_n：每一类中允许的最少模式数目
  - θ_S：类内各分量分布的距离标准差上界(分裂用)
  - θ_D：两类中心间的最小距离下界(合并用)
  - L：在每次迭代中可以合并的类的最多对数
  - I：允许的最多迭代次数

ISODATA算法

- ISODATA聚类分析的基本步骤
  1. 选择参数
  2. 确定初始聚类中心
  3. 用K均值算法对样本进行聚类
  4. 合并/分裂
  5. 计算各类的新的聚类中心
  6. 判断是否满足结束条件，如果不满足则转步骤3

延伸阅读

- 度量学习（Distance Metric Learning）
  - ECCV2010 Tutorial: http://www.cs.huji.ac.il/~ofirpele/DFML_ECCV2010_tutorial/
  - Matlab Toolbox: http://www.cs.cmu.edu/~liuy/distlearn.htm

- 数据聚类的新进展
  - 基于消息传递的数据聚类: http://www.psi.toronto.edu/affinitypropagation/FreyDueckScience07.pdf
  - Prof. Anil Jain的演讲: http://v.youku.com/v_show/id_XNDk4NDc3MDcy.html

本回回顾

- 聚类和分类的区别
- 聚类算法的基本要素
- 典型的聚类算法
  - K-均值聚类、ISODATA

下回预告

- 统计学习理论、支持向量机
  - 统计分析
  - 小样本问题
  - 线性支持向量机

- 准备好了吗？
  - 样本、统计分析的基本步骤
  - 分类超平面
  - 优化模型