模式识别

主讲：崔林艳&邹征夏

单位：宇航学院

开课专业：飞行器控制与信息工程

第六章 非监督分类器

CONTENTS PAGE

6.1 引言

【监督】

对于大多数统计模式识别问题，如果训练样本集中的每个样本的所属类别是已知，称为样本带标签，对应的模式识别问题称为监督问题，对应的模式识别方法称为监督方法。

【非监督】

对于某些统计模式识别问题，如果训练样本集中的每个样本的所属类别是未知的，称为样本不带标签，对应的模式识别问题称为非监督问题，对应的模式识别方法称为非监督方法。

聚类是最常见的无监督方法。

6.1 引言

【为什么要研究非监督学习问题？】

获取带标签的训练样本集的难度大

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

6.1 引言

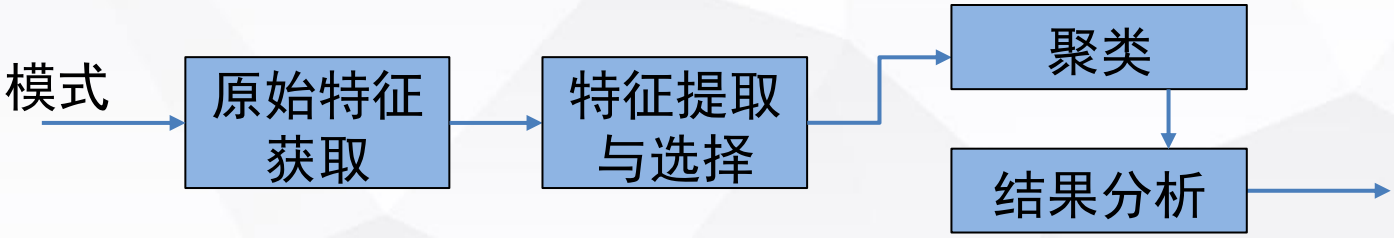
- 无监督模式识别的过程：

① 原始特征获取。对样本进行观测和预处理，获得样本的原始特征。

② 特征提取与选择。采用一定的算法对原始特征进行提取和选择。

③ 聚类。选择相关的无监督模式识别方法，利用样本进行聚类分析。

④ 结果分析。对聚类的结果进行评价；对聚类结果的合理性进行分析和解释；

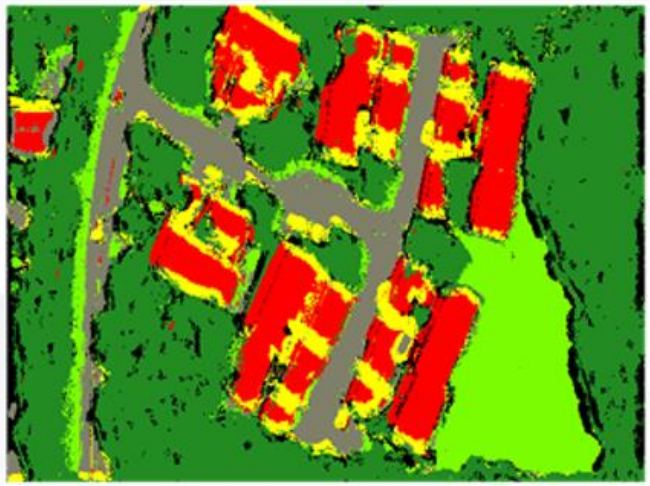


6.1 引言

例：多光谱遥感图像的无监督分类

遥感图像分类的目的是区分图像中不同种类的地物，具有同种特性的地物集合称为一类，一类地物具有同一标志，不同种类地物具有不同的光谱特性。

城市遥感图像

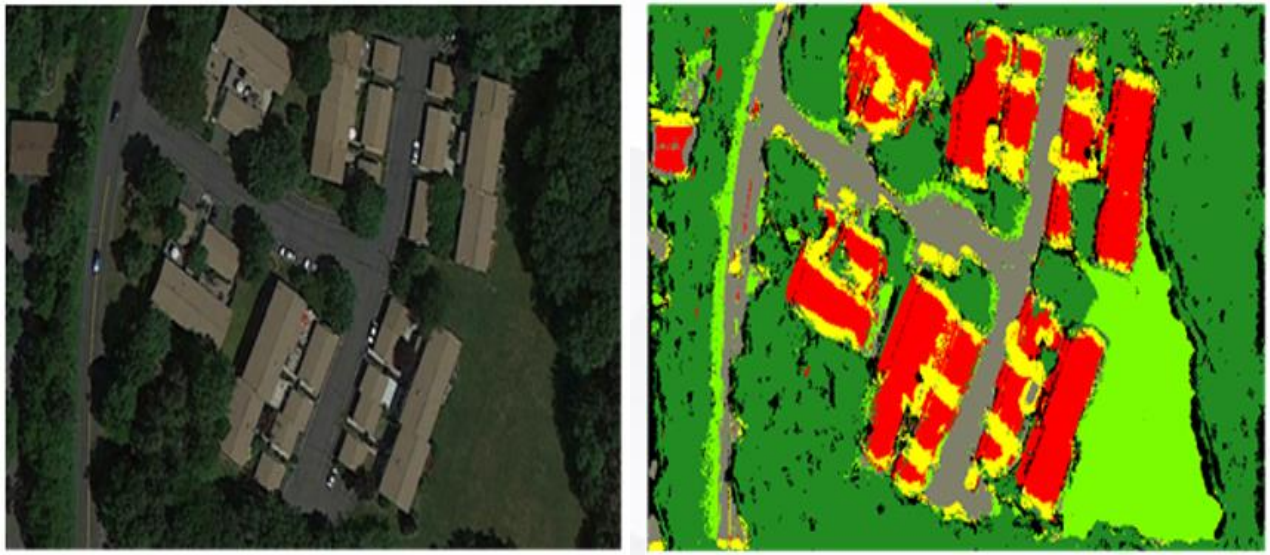


聚类结果

6.1 引言

例：多光谱遥感图像的无监督分类

1. 确定地物类别数；
2. 选择相似性测度；
3. 利用聚类算法将像素分割为若干具有相同光谱特性的集合；
4. 对分割后的结果进行统计分布特性评估。



6.2 聚类原理

6.2.1 聚类分析

6.2.2 相似性测度

6.2.3 聚类准则函数

6.2.4 聚类方法

6.2.5 小结

6.2 聚类原理

6.2.1 聚类分析

一种常用的非监督学习方法，该方法基于样本之间的相似和靠近程度，对无标签的样本进行逐步聚合（称为聚类 **Clustering**），实现子集的划分。

- 聚类的过程:
  - 迭代算法
- 假设条件：
  - 同类样本相互“靠近”
  - 异类样本相互“远离”

6.2 聚类原理

6.2.2 相似性测度

聚类过程基于样本之间的相似和靠近程度来逐步进行，因此，用来衡量样本间相似性程度的数学度量或测度，称为相似性测度。

利用样本间的相似性测度值对样本进行归类或子集划分，从而实现聚类。

- （1）欧氏距离
- （2）明氏距离
- （3）马氏距离
- （4）互相关函数

6.2 聚类原理

6. 2. 2 相似性测度

设$\mathbf{x}$ 和 $\mathbf{x}'$为两个样本的特征向量，d为样本空间维数

➢ （1）欧氏距离

最常见的两点之间或多点之间的距离表示方法，**在聚类中多数采用该距离测度**。又称为欧几里得度量（Euclidean Metric），定义于欧几里得空间中。

$$D = \| \mathbf{x} - \mathbf{x}' \|_2 = \sqrt{\sum_{i=1}^{d} \left| x_i - x_i' \right|^2}$$

问题分析：将样本的不同属性（即各变量）之间的差别等同看待，未区分不同属性差异。实际应用中可通过特征归一化方式，减少该问题。

6.2.2 相似性测度

设$\mathbf{x}$和$\mathbf{x}'$为两个样本的特征向量，d为样本空间维数

- **（2）明氏距离**

$$M'(\mathbf{x},\mathbf{x}')=\left(\sum_{i=1}^{d}\left|x_i-x_i'\right|^q\right)^{1/q}$$

其中$q\geq 1$

a) 当 q=1时，明氏距离等价于曼哈顿距离。

$$M'(\mathbf{x},\mathbf{x}')=\sum_{i=1}^{d}\left|x_i-x_i'\right|$$

b) 当q=2时，明氏距离等价于欧氏距离。

c) 当 $q=\infty$ 时，明氏距离等价于切比雪夫距离。

$$M'(\mathbf{x},\mathbf{x}')=\max_{1\leq i\leq d}\left|x_i-x_i'\right|$$

6.2 聚类原理

6.2.2 相似性测度

设$\mathbf{x}$和$\mathbf{x}'$为两个样本的特征向量，d为样本空间维数

➢ （3）马氏距离

马氏距离（Mahalanobis Distance）是由马哈拉诺比斯（P. C. Mahalanobis）提出的，**表示数据的协方差距离**。它是一种有效的计算两个未知样本相似度的方法。

$$M(\mathbf{x},\mathbf{x}')=\left[(\mathbf{x}-\mathbf{x}')^T\sum{}^{-1}(\mathbf{x}-\mathbf{x}')\right]^{\frac{1}{2}}$$

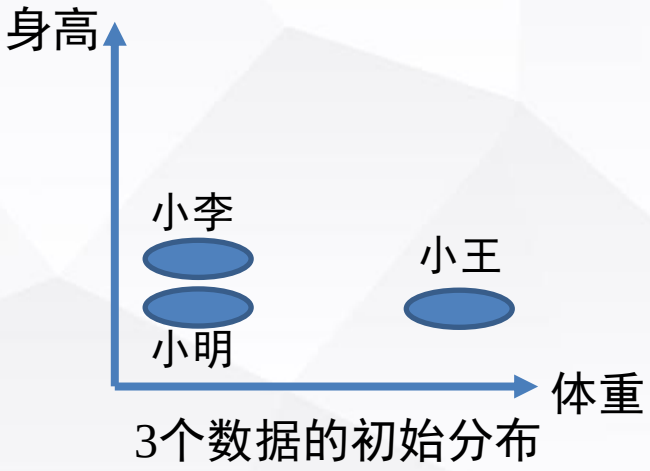
Σ为样本的协方差矩阵。当协方差矩阵为单位矩阵（各特征之间的相关程度一样）时，马氏距离简化为欧氏距离；当协方差矩阵为对角阵（特征只与自己相关，与其他特征无关）时，各特征完全相互独立。

6.2 聚类原理

- 举例：如果以厘米为单位来测量人的身高，以克为单位测量人的体重。每个人被表示为一个两维向量，如一个人身高173cm，体重50000g（50kg），表示为（173,50000），根据身高体重的信息来判断体型的相似程度。

已知小明（160,60000）；小王（160,59000）；小李（170，60000）。

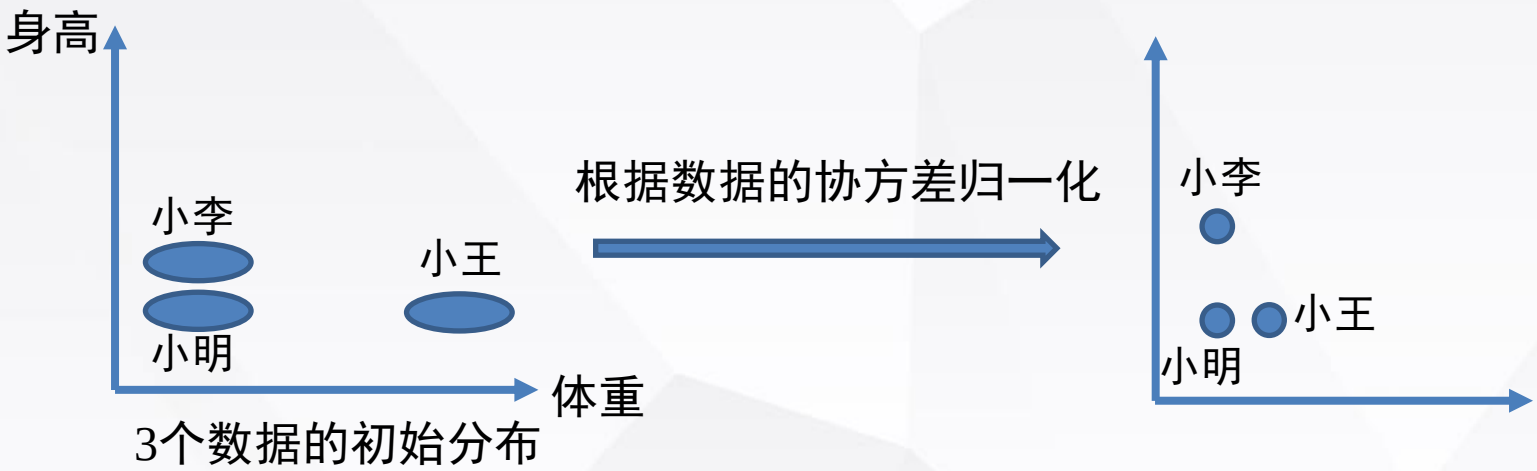
利用欧氏距离判断：小明与小李体型相似。(与常识不符)



3个数据的初始分布

6.2 聚类原理

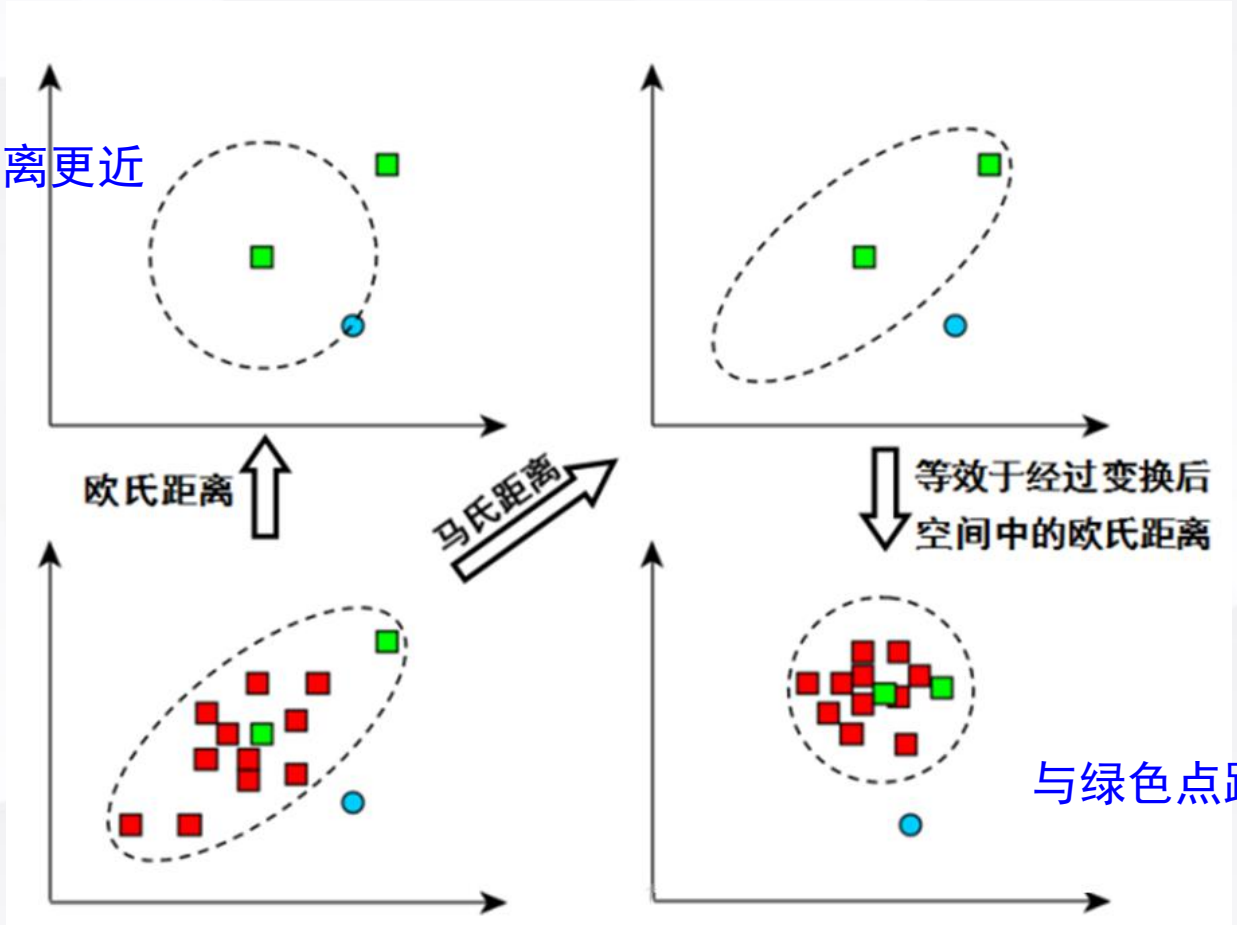
- 马氏距离通过除以协方差矩阵，将各个分量之间的方差除掉，消除了量纲性，与实际情况更吻合。



$$M(\mathbf{x},\mathbf{x}') = \left[(\mathbf{x}-\mathbf{x}')^T \sum{}^{-1} (\mathbf{x}-\mathbf{x}')\right]^{\frac{1}{2}}$$

6.2 聚类原理

- 欧氏距离VS马氏距离



6. 2. 2 相似性测度

➢ （4）互相关函数

a) 角度相关系数

$$\cos(\mathbf{x},\mathbf{x}') = \frac{\mathbf{x}\cdot\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|} = \frac{\sum_{i=1}^{d} x_i x_i'}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d}\left(x_i'\right)^2}}$$

表征两个向量的夹角余弦。两个向量越相似，该值越接近1，越不相似，该值越接近-1.

6.2.2 相似性测度

➢ （4）互相关函数

b) 皮尔逊相关系数

$$\rho(\mathbf{x},\mathbf{x}') = \frac{Cov(\mathbf{x},\mathbf{x}')}{\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'}}$$

- 两个向量之间的协方差除以它们各自标准差的乘积。系数的取值总是在-1.0到1.0之间。
- 当相关系数的绝对值越大，相关性越强。当相关系数为0时，两向量之间不相关。

6.2 聚类原理

6.2.1 聚类分析

6.2.2 相似性测度

6.2.3 聚类准则函数

6.2.4 聚类方法

6.2.5 小结

6.2 聚类原理

6.2.3 聚类准则函数

设有样本集合 $\boldsymbol{Z}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$;
要划分成C个不相交的子集$\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots\boldsymbol{Z}_C$

- **准则1：误差平方和准则**（简单且应用广泛的准则函数）

令 n_i 表示子集 Z_i 中样本的数量，$\boldsymbol{m}_i$表示子集Z_i 样本的均值向量

$$\boldsymbol{m}_i=\frac{1}{n_i}\sum_{\boldsymbol{x}\in\boldsymbol{Z}_i}\boldsymbol{x}$$

误差平方和准则定义为：

$$J_e=\sum_{i=1}^{C}\sum_{\boldsymbol{x}\in\boldsymbol{Z}_i}\|\boldsymbol{x}-\boldsymbol{m}_i\|_2^2$$

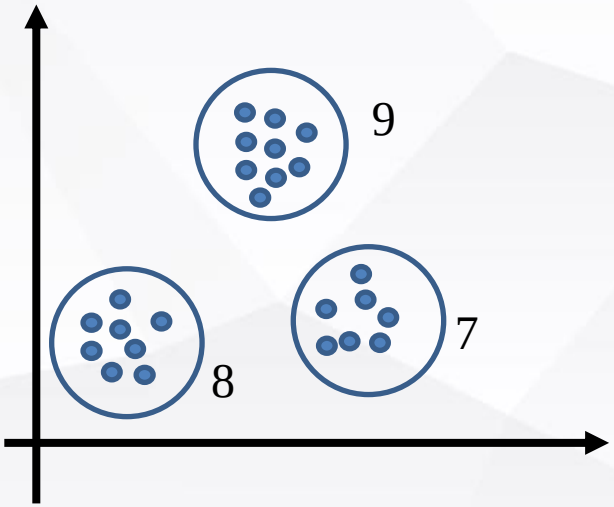
一个好的聚类方法应能使每个子集中的所有向量与这个均值向量的“误差向量”长度平方之和最小。

6.2.3 聚类准则函数

➢ 准则1：误差平方和准则

J_e 的值取决于类别的数目和样本的分布情况，使得该值最小的划分应该是最优的聚类，或称为最小方差聚类。

- 误差平方和准则适用于各类样本比较密集且样本数目差异不大的情况：



共有3个类型。各个类型的样本数目差异不大（分别为7，8，9个）。类内较密集，误差平方和很小，类别之间距离远。

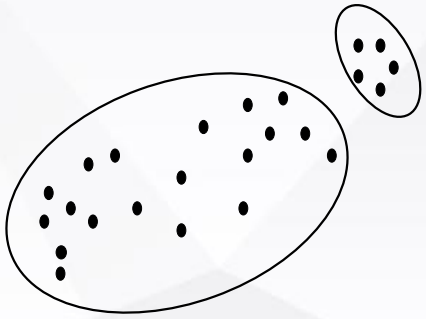
6.2 聚类原理

6.2.3 聚类准则函数

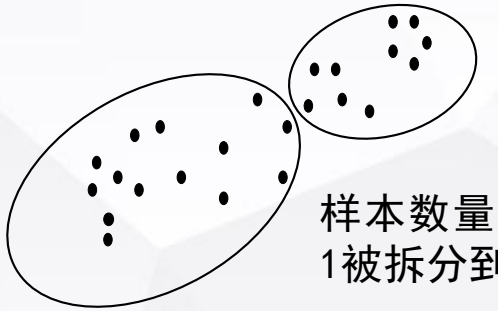
➢ 准则1：误差平方和准则

J_e 的值取决于类别的数目和样本的分布情况，使得该值最小的划分应该是最优的聚类，或称为最小方差聚类。

- 当不同类型的样本数目相差很大，如果采用误差平方和准则，有可能把样本数目多的类型分开，以使得总的 J_e 最小：



合理聚类，但J_e值较大



不合理聚类，但J_e值较小

6.2.3 聚类准则函数

➢ 准则2：离散度准则

基于类内离散度矩阵、类间离散度矩阵和总离散度矩阵，所定义的聚类准则。

离散度矩阵

① 聚类Z_i的离散度矩阵为：$$S_i = \sum_{x \in Z_i} (x - m_i)(x - m_i)^T$$

② 总的类内离散度矩阵为：$$S_w = \sum_{i=1}^{C} S_i$$

③ 类间离散度矩阵为：$$S_b = \sum_{i=1}^{C} (m_i - m)(m_i - m)^T$$

m是样本总均值向量

④ 全部样本的总离散度矩阵为：

$$S_t = S_w + S_b$$

6.2 聚类原理

6.2.3 聚类准则函数

➢ 准则2：离散度准则

① 离散度矩阵迹准则：$J_1 = tr\left(S_w^{-1}S_b\right)$　　$J_2 = tr\left(S_w^{-1}S_t\right)$

*tr*表示矩阵的迹，即矩阵对角线元素之和

② 离散度矩阵行列式准则：$J_3 = \left|S_w^{-1}S_b\right|$　　$J_4 = \left|S_w^{-1}S_t\right|$

$|.|$表示行列式

以上准则同时考虑了类内散度和类间散度。为了得到好的聚类结果，它们的取值越大越好。

6.2 聚类原理

6.2.1 聚类分析

6.2.2 相似性测度

6.2.3 聚类准则函数

6.2.4 聚类方法

6.2.5 小结

6.2.4 聚类方法

① 基于划分的聚类方法

- 是一种自顶向下的方法，对于给定的由N个样本数据组成的数据集Z，将该数据集划分为C个分区，其中每个分区代表一个簇。

> 给定要构建的分区数C，首先创建一个初始划分，然后采用迭代重定位技术，通过把对象从一个组移动到另一个组来改进划分。一个好的划分准则是：同一个簇中的相关对象尽可能相互“接近”或相关，而不同簇中的对象尽可能地“远离”或不同。

- 最经典的基于划分的聚类方法：K-均值聚类。指定聚类中心，再通过迭代的方式更新聚类中心。

6.2.4 聚类方法

② 基于层次的聚类方法

对给定的数据进行层次分解，直到满足某种条件为止。根据层次分解的顺序可分为凝聚和分裂的方法。

a) 自下而上的聚合法：将样本集中所有数据点当成初始聚类，共有N个聚类，第一步将相似性最强的两个类别合并，得到N-1个聚类，以此类推直到预期的聚类数目。（N类-》C类的过程）

b) 自上而下的分裂法：首先将整个数据集看成一类，然后递归地进行划分，直至所有数据点都被分在了不同聚类中。（1类-》2类-》…-》C类的过程）

6.2.4 聚类方法

③ 基于密度的聚类方法

- 基于距离的聚类算法一般只能发现球状簇，不适用于具有任意形状的簇。
- 基于密度的聚类方法从数据分布区域的密度着手，核心思想是寻找被低密度区域分离的高密度区域。通过连接密度较大的区域，形成不同形状的簇，可消除孤立点和噪声对聚类的影响，发现任意形状的簇。

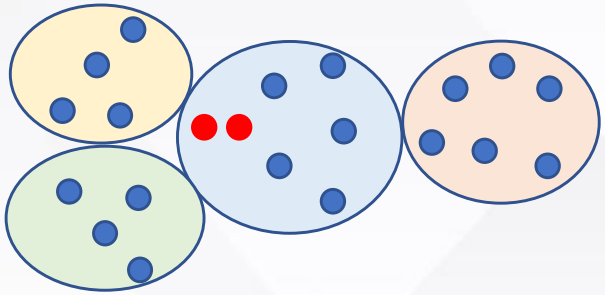
聚类中心的周围都是密度比其低的点，同时这些点距离该聚类中心的距离相比于其他聚类中心来说是最近的。

经典算法：DBSCAN

6.2.4 聚类方法

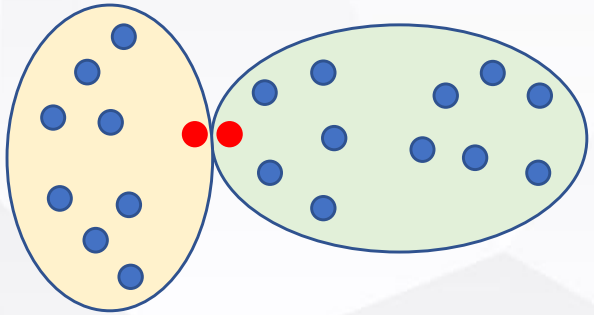
④ 基于模型的聚类方法

首先给每个簇假定一个模型（数据在空间分布中的密度函数或者其他函数），然后寻找能够很好满足该模型的数据集。该聚类方法假定目标数据集是由一系列的概率分布所决定的。



基于距离的聚类结果

（将距离近的点聚集在一起）



基于模型的聚类结果

（按照特定的概率分布模型进行聚类）

6.2 聚类原理

6.2.5 小结

- 相似性测度

  基于距离的测度：欧氏距离、马氏距离、明氏距离

  基于相似性函数的测度：角度相关函数、皮尔逊相关系数

- 聚类准则函数

  误差平方和准则：最常用的准则

  离散度准则：离散度矩阵迹准则、离散度矩阵行列式准则

- 聚类方法分类

  基于划分的聚类方法（K-均值）、基于层次的聚类方法（分级聚类）

  基于密度的聚类方法（DBSCAN）、基于模型的聚类方法

第六章 非监督分类器

CONTENTS PAGE

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

6.3.2 算例分析

6.3.3 分级聚类小结

6.3.4 作业

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

- 分级聚类，也称为层次聚类（ Hierarchical Clustering），是最常见的聚类分析算法。

- 给定 d 维特征空间中由 N 个数据点所构成的样本集

$$\boldsymbol{D}=\left\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\right\}$$

- 聚类 $\Omega=\left\{\omega_1,\omega_2,...,\omega_C\right\}$ 将数据集划分成 C个类别，假设类别与类别之间没有重合的数据点。

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

➢ 实现步骤

① 初始化：首先将每个数据点都看成一类，

$$\Omega = \{\omega_1, \omega_2, ..., \omega_N\}, \omega_i = \{\boldsymbol{x}_i\}$$

② 聚合：计算所有类别之间的相似性测度，将最相似的两个类别合并。

③ 更新：重复上述过程直至达到预期的类别个数。

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

➢ 类别间相似性度量

两个类别之间的相似性测度通常采用数据点之间的欧氏距离

$$d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{d}|x_i-y_i|^2}$$

（1）最小距离法

对于类别ω_i和ω_j，将两类别中最靠近的两个数据点作为类别的代表，将类间相似性测度定义为类别ω_i和ω_j最靠近的这两个数据点之间的欧氏距离：

$$d(\omega_i,\omega_j)=\min\left\{d(\boldsymbol{x},\boldsymbol{y})\,|\,\boldsymbol{x}\in\omega_i,\boldsymbol{y}\in\omega_j\right\}$$

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

- 类别间相似性度量

两个类别之间的相似性测度通常采用数据点之间的欧氏距离

$$d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{d}|x_i-y_i|^2}$$

（2）最大距离法

对于类别ω_i和ω_j，将两类别中最远的两个数据点作为类别的代表，将类间相似性测度定义为类别ω_i和ω_j相距最远两个数据点之间的距离：

$$d(\omega_i,\omega_j)=\max\left\{d(\boldsymbol{x},\boldsymbol{y})\,|\,\boldsymbol{x}\in\omega_i,\boldsymbol{y}\in\omega_j\right\}$$

6.3 分级聚类

6.3.1 基于聚合法的分级聚类

➢ 类别间相似性度量

两个类别之间的相似性测度通常采用数据点之间的欧氏距离

$$d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{d}|x_i-y_i|^2}$$

（3）均值距离法

将两个类别的均值点作为类别的代表，将均值点之间的距离作为类别相似性测度：

$$d(\omega_i,\omega_j)=d(\boldsymbol{m}_i,\boldsymbol{m}_j)$$

其中 $\boldsymbol{m}_i=\frac{1}{N_i}\sum_{\boldsymbol{x}\in\omega_i}\boldsymbol{x}$

N_i 表示类别 ω_i 中数据点的个数

6.3.1 基于聚合法的分级聚类

➢ 类别间相似性度量

两个类别之间的相似性测度通常采用数据点之间的欧氏距离

$$d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{d}|x_i-y_i|^2}$$

（4）组平均距离法

将两个类别之间所有点对距离的平均值作为类别相似性测度。相比于最小距离法和最大距离法，组平均距离法可以有效避免类内特殊点带来的不利影响。

$$d(\omega_i,\omega_j)=\frac{\sum_{\mathbf{x}\in\omega_i}\sum_{\mathbf{y}\in\omega_j}d(\boldsymbol{x},\boldsymbol{y})}{N_i\cdot N_j}$$

6.3.1 基于聚合法的分级聚类

➢ 类别间相似性度量

（5）沃德距离法

又称最小方差法，将两个类别之间的相似性测度定义为两个类别合并时的平方和误差（SSE）的增量。

- 给定一个聚类ω_i，其SSE定义为：

$$\mathrm{SSE}_i = \sum_{\boldsymbol{x}\in\omega_i} \|\boldsymbol{x}-\boldsymbol{m}_i\|_2^2$$

- 当沃德距离将ω_i和ω_j合并成ω_{ij}时，将SSE值的变化量作为两个类别之间的相似性测度：

$$d(\omega_i,\omega_j) = SSE_{ij} - SSE_i - SSE_j$$

6.3 分级聚类

6.3.2 算例分析

假设有一无标签样本集共有5个数据点A、B、C、D、E：A = [1 3 3]^T, B = [2 0 4]^T, C = [3 2 2]^T, D = [0 2 3]^T, E = [1 2 3]^T，现要求利用最小距离法将该样本集划分为两类。

基于聚合法的分级聚类

① 初始化：首先将每个数据点都看成一类，

$$\Omega = \{\omega_1, \omega_2, ..., \omega_N\}, \omega_i = \{\boldsymbol{x}_i\}$$

② 聚合：计算所有类别之间相似性测度，将最相似的两个类别合并。

③ 更新：重复上述过程直至达到预期的类别个数。

6.3 分级聚类

6.3.2 算例分析

$A=[1\ \ 3\ \ 3]^T$, $B=[2\ \ 0\ \ 4]^T$, $C=[3\ \ 2\ \ 2]^T$,
$D=[0\ \ 2\ \ 3]^T$, $E=[1\ \ 2\ \ 3]^T$

➢ 求解

起初每个数据点自成一类，即聚类成：

$$\omega_1=\{A\},\omega_2=\{B\},\omega_3=\{C\},\omega_4=\{D\},\omega_5=\{E\}$$

计算类别之间的最小距离，并将类别之间的最小距离列在表中：

$d(\omega_i,\omega_j)$	$\omega_2=\{B\}$	$\omega_3=\{C\}$	$\omega_4=\{D\}$	$\omega_5=\{E\}$
$\omega_1=\{A\}$	$\sqrt{11}$	$\sqrt{6}$	$\sqrt{2}$	1
$\omega_2=\{B\}$		3	3	$\sqrt{6}$
$\omega_3=\{C\}$			$\sqrt{10}$	$\sqrt{5}$
$\omega_4=\{D\}$				1

合并 ω_1 和 ω_5

6.3.2 算例分析

$A=[1\ \ 3\ \ 3]^T$, $B=[2\ \ 0\ \ 4]^T$, $C=[3\ \ 2\ \ 2]^T$,
$D=[0\ \ 2\ \ 3]^T$, $E=[1\ \ 2\ \ 3]^T$

➢ 求解

得到新的聚类结果：$\omega_1=\{A,E\}$ $\omega_2=\{B\}$ $\omega_3=\{C\}$ $\omega_4=\{D\}$

利用最小距离法，重新计算类间最小距离。以 $d(\omega_1,\omega_2)$ 为例

$$d(\omega_1,\omega_2)=\min\{d(A,B),d(E,B)\}=\sqrt{6}$$

同理可依次计算其他类间距离，得到新的类间最小距离：

$d(\omega_i,\omega_j)$	$\omega_2=\{B\}$	$\omega_3=\{C\}$	$\omega_4=\{D\}$
$\omega_1=\{A,E\}$	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{1}$
$\omega_2=\{B\}$		3	3
$\omega_3=\{C\}$			$\sqrt{10}$

合并 ω_1 和 ω_4

6.3.2 算例分析

➢ 求解

得到新的聚类结果：$\omega_1=\{A,E,D\}$ $\omega_2=\{B\}$ $\omega_3=\{C\}$

利用最小距离法，重新计算类间最小距离，得到新的类间最小距离：

$d(\omega_i,\omega_j)$	$\omega_2=\{B\}$	$\omega_3=\{C\}$
$\omega_1=\{A,E,D\}$	$\sqrt{6}$	$\sqrt{5}$
$\omega_2=\{B\}$		3

合并 ω_1 和 ω_3

得到新的聚类结果：$\omega_1=\{A,E,D,C\}$ $\omega_2=\{B\}$

此时聚成两类，聚类过程结束。

6.3.3 分级聚类小结

- 分级聚类最简单的聚类方法
  一遍迭代完成，$N \rightarrow C$。

- 原理简单易实现
  一次合并掉一个聚类。

- 相似性测度采用欧氏距离
  - “靠近”和“远离”是真实意义的；
  - 采用不同距离计算方法，可能会导致结果差异。

6.3.4 作业

针对分级聚类算例，假设有一无标签样本集共有5个数据点A、B、C、D、E：$A = [1\ \ 3\ \ 3]^T$，$B = [2\ \ 0\ \ 4]^T$，$C = [3\ \ 2\ \ 2]^T$，$D = [0\ \ 2\ \ 3]^T$，$E = [1\ \ 2\ \ 3]^T$，现要求利用最大距离法、均值距离法将该样本集划分为两类。

第六章 非监督分类器

CONTENTS PAGE

6.4 动态聚类

6.4.1 动态聚类原理

6.4.2 K-均值聚类

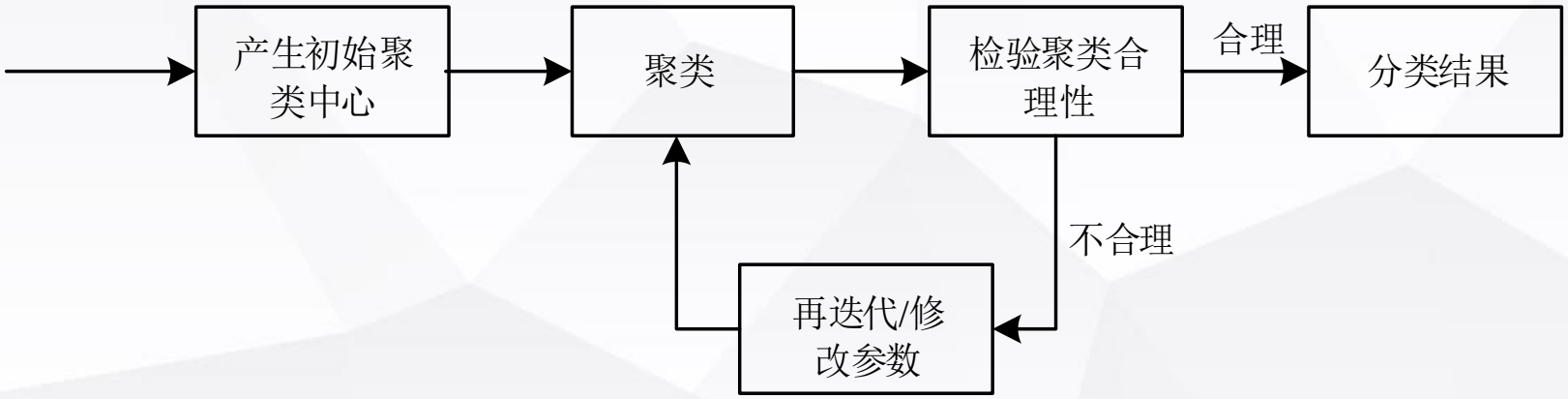
6.4.3 ISODATA算法

6.4.4 小结

6.4 动态聚类

6.4.1 动态聚类原理

- 基本思路：首先选择一批有代表性的样本作为初始聚类中心，将样本集进行初始分类；然后根据聚类准则重新计算聚类中心，不断调整不合适的聚类样本，进行重新聚类，直到满足给定的结束条件为止。（从初始的随意性划分到“最优”划分，是一个动态迭代过程）



动态聚类的基本流程图

6.4 动态聚类

6.4.2 K-均值聚类

又称“C-均值算法”，算法的基础是误差平方和准则。其基本思想是，通过迭代寻找C个聚类的划分方案，使得用这C个聚类的均值来表达相应各类样本时所得到的总体误差最小。

- 若 N_i 是第 i 聚类 ω_i 中的样本数目，$\boldsymbol{m}_i$ 是这些样本的均值，即

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{\boldsymbol{x}\in\omega_i}\boldsymbol{x}$$

- 把 ω_i 中的各样本$\boldsymbol{x}$与均值 $\boldsymbol{m}_i$ 间的误差平方和对所有类相加：

$$J_e = \sum_{i=1}^{C}\sum_{\boldsymbol{x}\in\omega_i} \| \boldsymbol{x}-\boldsymbol{m}_i \|_2^2$$

J_e 是误差平方和聚类准则

6.4 动态聚类

6.4.2 K-均值聚类

$$J_e = \sum_{i=1}^{C} \sum_{\boldsymbol{x} \in \omega_i} \| \boldsymbol{x} - \boldsymbol{m}_i \|_2^2$$

- 该公式一定程度上刻画了类内样本围绕类均值向量的紧密程度，J_e越小则类内样本相似度越高。
- 对于不同的聚类，使J_e极小的聚类是误差平方和准则下的最优结果。
- 极小化J_e。K-均值算法采用迭代优化近似求解。

6.4.2 K-均值聚类

➢ 分析，把样本x从ω_k 类移入ω_j 类对误差平方和的影响：

① 设从 ω_k 中移出 $\boldsymbol{x}$ 后的集合为$\tilde{\omega}_k$ ，它相应的均值是 $\tilde{\boldsymbol{m}}_k$

$$\tilde{\boldsymbol{m}}_k = \boldsymbol{m}_k + \frac{1}{N_k - 1}\left(\boldsymbol{m}_k - \boldsymbol{x}\right)$$

式中的 $\boldsymbol{m}_k$ 和 N_k 是 ω_k 的样本均值和样本数.

② 设 ω_j 接受 $\boldsymbol{x}$ 后的集合为 $\tilde{\omega}_j$, 它相应的均值是 $\tilde{\boldsymbol{m}}_j$

$$\tilde{\boldsymbol{m}}_j = \boldsymbol{m}_j + \frac{1}{N_j + 1}\left(\boldsymbol{x} - \boldsymbol{m}_j\right)$$

式中的 $\boldsymbol{m}_j$ 和 N_j 是 ω_j 的样本均值和样本数.

6.4.2 K-均值聚类

➢ 分析，把样本x从ω_k 类移入ω_j 类对误差平方和的影响：

③ x的移动只影响 ω_k和ω_j 两类，对其他类无任何影响，因此只需要计算这两类的新的误差平方和 $\tilde{J}_k$ 和 $\tilde{J}_j$

$$\tilde{J}_k = J_k - \frac{N_k}{N_k - 1}\left\|\boldsymbol{x} - \boldsymbol{m}_k\right\|_2^2 \qquad \tilde{J}_j = J_j + \frac{N_j}{N_j + 1}\left\|\boldsymbol{x} - \boldsymbol{m}_j\right\|_2^2$$

如果

$$\frac{N_j}{N_j + 1}\left\|\boldsymbol{x} - \boldsymbol{m}_j\right\|_2^2 < \frac{N_k}{N_k - 1}\left\|\boldsymbol{x} - \boldsymbol{m}_k\right\|_2^2$$

则把样本$\boldsymbol{x}$ 从ω_k 移入到ω_j 就会使总误差平方和减少。只有当$\boldsymbol{x}$ 离 $\boldsymbol{m}_j$的距离比离 $\boldsymbol{m}_k$的距离更近时才满足上述不等式。

6.4 动态聚类

6.4.2 K-均值聚类

K-均值算法通过将样本不断重新划分，从而调整相应聚类中心，优化聚类结果，最终使得误差平方和极小。

- 基本流程：假设聚成C类，则：

Step1：确定C个初始聚类，计算相应的聚类中心 $\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_C$

Step2：选择一个备选样本 $\boldsymbol{x}$，设 $\boldsymbol{x}$ 现在在 ω_i 中

Step3：若 $N_i = 1$,则转2,否则继续

Step4：计算将样本$\boldsymbol{x}$移入到第j个类别所引起的误差平方和变化

$$d_j = \begin{cases} \dfrac{N_j}{N_j + 1} \left\| \boldsymbol{x} - \boldsymbol{m}_j \right\|_2^2 & j \neq i \\ \dfrac{N_i}{N_i - 1} \left\| \boldsymbol{x} - \boldsymbol{m}_i \right\|_2^2 & j = i \end{cases} \Rightarrow J_j = \begin{cases} J_j + \dfrac{N_j}{N_j + 1} \left\| \boldsymbol{x} - \boldsymbol{m}_j \right\|_2^2 & j \neq i \uparrow \\ J_j - \dfrac{N_i}{N_i - 1} \left\| \boldsymbol{x} - \boldsymbol{m}_i \right\|_2^2 & j = i \downarrow \end{cases}$$

6.4.2 K-均值聚类

Step5： 若样本x到第k个类别聚类中心所产生的误差平方和变化最小，则把样本x从ω_i移到ω_k中。

Step6： 重新计算$\boldsymbol{m}_i$ 和$\boldsymbol{m}_k$的值,并修改误差平方和J_e

$$J_e = \sum_{i=1}^{C} \sum_{\boldsymbol{x} \in \omega_i} \| \boldsymbol{x} - \boldsymbol{m}_i \|_2^2$$

Step7： 反复迭代，直至所有样本均不能移动，J_e不发生变化。

以上聚类算法通过不断计算和修改聚类准则函数（误差平方和J_e），会导致计算量较大，为进一步优化算法可采用更为简单的计算方法。

6.4 动态聚类

6.4.2 K-均值聚类

➢ 基本流程优化版：

Step1：任意选取C个样本作为初始的聚类中心 $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_C$

Step2：遍历所有待分类的样本$\boldsymbol{x}_i$，计算该样本与每个聚类中心样本的距离，按照最小距离原则将样本划分到C类中的某一类；

Step3：计算新分类后的聚类中心；

Step4：如果任意一个类的聚类中心都不再发生变化，则结束，否则转步骤2；

上述算法中，通过计算和修改聚类中心，没有直接运用聚类准则函数进行分类，减少了运算。

6.4.2 K-均值聚类

- **关键问题1：样本集初始划分**

第一步：代表点的选择

a) 方式1：凭经验选择代表点，从数据中找出从直观上看来是比较合适的代表点；

b) 方式2：将全部数据随机地分成C类，计算每类重心，将这些重心作为每类代表点；

c) 方式3：用前C个样本点作为代表点。

6.4.2 K-均值聚类

➢ 关键问题1：样本集初始划分

第二步：确定代表点后进行初始分类

a) **初始划分方式1**：将其余样本点按照与代表点距离最近原则，划分到相应类中（代表点保持不变）；

b) **初始划分方式2**：每个代表点自成一类，将样本依顺序归入与其最近的代表点那一类，并立即重新计算该类的重心以代替原来的代表点（代表点进行动态更新）。然后再计算下一个样本的归类，直至所有的样本都归到相应的类为止。

6.4.2 K-均值聚类

- **关键问题2：类别间相似性测度**

类别间相似性测度的选择影响分类的结果。

a） **欧氏距离**

b） 明氏距离

c） 马氏距离

d） 相关性测度

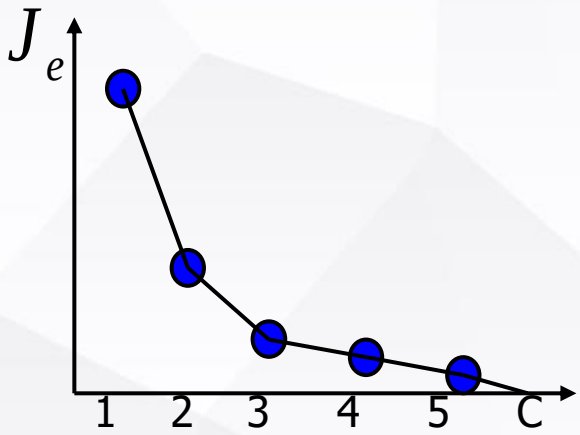
6.4.2 K-均值聚类

- **关键问题3：类别数的确定**

a） 方式1：根据经验人为确定类别数；

b） 方式2：通过算法自动产生

例如：绘制Je-C曲线、聚类有效性评价函数



聚类准则函数Je随类别数C的变化曲线

6.4.2 K-均值聚类

➢ **关键问题3：类别数的确定**

聚类有效性评价函数。到目前为止，已提出了多种聚类有效性标准，其共同目标是使分类结果达到类内紧密、类间远离。

以SD有效性函数为例，于2000年提出，是基于聚类平均散布性和聚类间总体分离性的一种相对度量方法。

① 聚类平均散布性：

$$Scat(C)=\frac{1}{C}\sum_{i=1}^{C}\|\sigma(\omega_i)\|/\|\sigma(\Omega)\|$$

$\sigma(\Omega)$：样本集Ω的方差

$\sigma(\omega_i)$：聚类ω_i的方差

6.4.2 K-均值聚类

➢ **关键问题3：类别数的确定**

② 聚类间总体分离性：

$$Dis(C)=\frac{D_{\max}}{D_{\min}}\sum_{j=1}^{C}\left(\sum_{i=1}^{C}\left\|\boldsymbol{m}_i-\boldsymbol{m}_j\right\|\right)^{-1}$$

$D_{\max}$和$D_{\min}$ 分别表示聚类中心间的最大和最小距离：

$$D_{\max}=\max\left(\left\|\boldsymbol{m}_i-\boldsymbol{m}_j\right\|\right)\qquad D_{\min}=\min\left(\left\|\boldsymbol{m}_i-\boldsymbol{m}_j\right\|\right)$$

③ 聚类有效性函数：

$$SD(C)=\alpha Scat(C)+Dis(C)$$

具有最小SD值所对应的C即是最佳的类别数

6.4.2 K-均值聚类

- **关键问题3：类别数的确定**

类别数的确定问题是K-均值目前仍然无法完美解决的问题，一般需要具体问题具体分析。

6.4.2 K-均值聚类

➢ 算例分析：

已知无标签样本集：

$$\boldsymbol{x}_1=\begin{bmatrix}0\\0\end{bmatrix}\quad \boldsymbol{x}_2=\begin{bmatrix}1\\1\end{bmatrix}\quad \boldsymbol{x}_3=\begin{bmatrix}1\\0.5\end{bmatrix}\quad \boldsymbol{x}_4=\begin{bmatrix}0.5\\1\end{bmatrix}\quad \boldsymbol{x}_5=\begin{bmatrix}2\\0.5\end{bmatrix}$$

$$\boldsymbol{x}_6=\begin{bmatrix}6\\2\end{bmatrix}\quad \boldsymbol{x}_7=\begin{bmatrix}7\\1\end{bmatrix}\quad \boldsymbol{x}_8=\begin{bmatrix}6\\1\end{bmatrix}\quad \boldsymbol{x}_9=\begin{bmatrix}7\\2\end{bmatrix}\quad \boldsymbol{x}_{10}=\begin{bmatrix}8\\3\end{bmatrix}$$

试用K-均值算法将这些样本聚成两类。

6.4 动态聚类

6.4.2 K-均值聚类

- 回顾K-均值聚类基本流程优化版：

Step1：任意选取C个样本作为初始的聚类中心 $\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_C$

Step2：遍历所有待分类的样本$\boldsymbol{x}_i$，计算该样本与每个聚类中心样本的距离，按照最小距离原则将样本划分到C类中的某一类；

Step3：计算新分类后的聚类中心；

Step4：如果任意一个类的聚类中心都不再发生变化，则结束，否则转步骤2；

6.4.2 K-均值聚类

➢ 算例求解：

① 初始代表点可以任意选取，在这里选

$$\boldsymbol{m}_1(0)=\boldsymbol{x}_1=\begin{bmatrix}0\\0\end{bmatrix}\qquad \boldsymbol{m}_2(0)=\boldsymbol{x}_2=\begin{bmatrix}1\\1\end{bmatrix}$$

② 因为：$\left\|\boldsymbol{x}_1-\boldsymbol{m}_1(0)\right\|_2<\left\|\boldsymbol{x}_1-\boldsymbol{m}_2(0)\right\|_2$ 所以 $\boldsymbol{x}_1\in\omega_1(1)$

$\left\|\boldsymbol{x}_2-\boldsymbol{m}_1(0)\right\|_2>\left\|\boldsymbol{x}_2-\boldsymbol{m}_2(0)\right\|_2$ 所以 $\boldsymbol{x}_2\in\omega_2(1)$

$\left\|\boldsymbol{x}_3-\boldsymbol{m}_1(0)\right\|_2>\left\|\boldsymbol{x}_3-\boldsymbol{m}_2(0)\right\|_2$ 所以 $\boldsymbol{x}_3\in\omega_2(1)$

。 。 。 。 。 。

得到：$\omega_1(1)=\{\boldsymbol{x}_1\},N_1=1;\qquad \omega_2(1)=\{\boldsymbol{x}_2,\boldsymbol{x}_3,\ldots,\boldsymbol{x}_{10}\},N_2=9;$

6.4.2 K-均值聚类

➢ 算例求解：

③ 计算新的聚类中心：

$$\boldsymbol{m}_1(1)=\frac{1}{N_1}\sum_{\boldsymbol{x}_i\in\omega_1(1)}\boldsymbol{x}_i=\boldsymbol{x}_1=\begin{bmatrix}0\\0\end{bmatrix}$$

$$\boldsymbol{m}_2(1)=\frac{1}{N_2}\sum_{\boldsymbol{x}_i\in\omega_2(1)}\boldsymbol{x}_i=\frac{1}{9}\left(\boldsymbol{x}_2+\boldsymbol{x}_3+\ldots+\boldsymbol{x}_{10}\right)=\begin{bmatrix}4.28\\1.33\end{bmatrix}$$

④ 因为 $\boldsymbol{m}_1(1)=\boldsymbol{m}_1(0)$ $\boldsymbol{m}_2(1)\neq\boldsymbol{m}_2(0)$ 所以转到(2)

⑤ 经过计算得到：

$$\omega_1(2)=\left\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5\right\},N_1=5$$

$$\omega_2(2)=\left\{\boldsymbol{x}_6,\boldsymbol{x}_7,\boldsymbol{x}_8,\boldsymbol{x}_9,\boldsymbol{x}_{10}\right\},N_2=5$$

6.4.2 K-均值聚类

➢ 算例求解：

⑥ 计算新的聚类中心：

$$\boldsymbol{m}_1(2)=\frac{1}{N_1}\sum_{\boldsymbol{x}_i\in\omega_1(2)}\boldsymbol{x}_i=\frac{1}{5}\left(\boldsymbol{x}_1+\boldsymbol{x}_2+\boldsymbol{x}_3+\boldsymbol{x}_4+\boldsymbol{x}_5\right)=\begin{bmatrix}0.9\\0.6\end{bmatrix}$$

$$\boldsymbol{m}_2(2)=\frac{1}{N_2}\sum_{\boldsymbol{x}_i\in\omega_2(2)}\boldsymbol{x}_i=\frac{1}{5}\left(\boldsymbol{x}_6+\boldsymbol{x}_7+\boldsymbol{x}_8+\boldsymbol{x}_9+\boldsymbol{x}_{10}\right)=\begin{bmatrix}6.8\\1.8\end{bmatrix}$$

⑦ 因为 $\boldsymbol{m}_1(2)\neq\boldsymbol{m}_1(1)$ $\boldsymbol{m}_2(2)\neq\boldsymbol{m}_2(1)$ ，所以转到(2)

⑧ 计算求得的分类结果与前一次的结果相同，即：

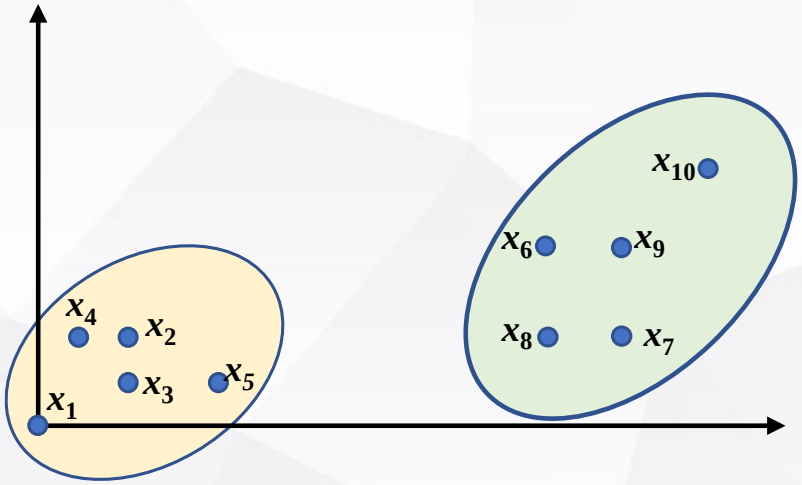
$$\omega_1(3)=\omega_1(2)\qquad\omega_2(3)=\omega_2(2)$$

6.4.2 K-均值聚类

➢ 算例求解：

⑨ 各聚类中心也与前一次的相同，$m_1(3)=m_1(2)$　$m_2(3)=m_2(2)$

因为 $m_1(3)=m_1(2)$　$m_2(3)=m_2(2)$，不会再出现新的类别划分，至此分类过程结束。



K-均值聚类后的分类结果

6.4.2 K-均值聚类

➢ **问题分析1：**

随机的初始中心选择对计算结果和迭代次数有较大的影响

改进方法： K-Means++

K-means++按照如下的思想选取C个聚类中心：

a) 在选取第一个聚类中心(n=1)时同样通过随机的方法。

b) 假设已经选取了n个初始聚类中心($0<n<C$)，则在选取第n+1个聚类中心时：距离当前n个聚类中心越远的点会有更高的概率被选为第n+1个聚类中心。

K-Means++初始聚类中心之间的相互距离要尽可能的远，可有效解决随机初始中心选择的问题。

6.4.2 K-均值聚类

- **问题分析2：**

聚类数量C需要预先设定

通过K-Means++可有效解决随机初始中心选择的问题，但是对于聚类数量的预先设定，在K-Means++中也没有很好的解决。

6.4 动态聚类

➢ 作业：

已知无标签样本集：

$$\boldsymbol{x}_1=\begin{bmatrix}1\\1\end{bmatrix}\quad \boldsymbol{x}_2=\begin{bmatrix}1\\2\end{bmatrix}\quad \boldsymbol{x}_3=\begin{bmatrix}2\\1\end{bmatrix}\quad \boldsymbol{x}_4=\begin{bmatrix}2\\2\end{bmatrix}\quad \boldsymbol{x}_5=\begin{bmatrix}2\\0\end{bmatrix}$$

$$\boldsymbol{x}_6=\begin{bmatrix}7\\1\end{bmatrix}\quad \boldsymbol{x}_7=\begin{bmatrix}8\\1\end{bmatrix}\quad \boldsymbol{x}_8=\begin{bmatrix}7\\2\end{bmatrix}\quad \boldsymbol{x}_9=\begin{bmatrix}8\\3\end{bmatrix}\quad \boldsymbol{x}_{10}=\begin{bmatrix}8\\0\end{bmatrix}$$

试用K-均值算法将这些样本聚成两类。

6.4 动态聚类

6.4.1 动态聚类原理

6.4.2 K-均值聚类

6.4.3 ISODATA算法

6.4.4 小结

6.4 动态聚类

6.4.3 ISODATA算法

➢ 算法思路：

Iterative **S**elf-**O**rganizing **D**ata **A**nalysis **T**echniques **A**lgorithm
（迭代自组织数据分析方法）

在K-均值算法基础上增加对聚类结果的“合并”和“分裂”：

- 聚类合并：当两个聚类中心之间距离值小于某个阈值时，将这两个聚类合并成一个；
- 聚类分裂：当某个聚类的样本方差（描述样本分散程度）大于一定的阈值且该聚类内样本数量超过一定阈值时，将该聚类分裂为两个聚类。

经过以上两个操作，可有效解决聚类数量需要设定的问题。

6.4.3 ISODATA算法

- ISODATA算法的输入：

① 预期的聚类中心数目C：

由用户指定一个聚类中心数目的参考值。ISODATA算法的聚类中心数目变动范围也由C决定，最终输出的聚类中心数目范围是 $[C/2, 2C]$。

② 每个类所包含的最少样本数目$N_{\min}$：

用于判断当某个类别所包含样本分散程度较大时是否可以进行分裂操作。如果分裂后会导致某个子类别所包含样本数目小于$N_{\min}$，就不会对该类别进行分裂操作。

6.4.3 ISODATA算法

- ISODATA算法的输入：

③ 最大方差Sigma：

用于衡量某个类别中样本的分散程度。当样本的分散程度超过这个值时，则有可能进行分裂操作（注意同时需要满足②中所述的条件）。

④ 两个类别对应的聚类中心之间所允许最小距离$d_{\min}$：

如果两个类别靠得非常近（即这两个类别对应聚类中心之间的距离非常小），则需要对这两个类别进行合并操作。是否进行合并的阈值由$d_{\min}$决定。

6.4 动态聚类

6.4.3 ISODATA算法

➢ 算法流程：

① 从数据集中随机选取C_0(可以不等于C)个样本作为初始聚类中心 $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_{C0}\}$；

② 针对数据集中每个样本$\boldsymbol{x}_i$，计算它到C_0个聚类中心的距离，并将该样本分到与其距离最小的聚类中心所对应的类中；

③ 判断上述每个类中的样本数量是否小于$N_{\min}$。如果小于$N_{\min}$，则需要删除该类，令$C_0=C_0-1$，并将该类中的样本重新分配到剩下与其距离最小的类中；

④ 针对每个类别，重新计算它的聚类中心；

6.4.3 ISODATA算法

➢ 算法流程：

⑤ 如果当前类别数 $\geq 2C$，则表明当前类别数太多，需进行合并操作；

⑥ 如果当前类别数$\leq C/2$，则表明当前类别数太少，需进行分裂操作；

⑦ 如果达到最大迭代次数则终止，否则回到第2步继续执行；

6.4.3 ISODATA算法

➢ 算法流程：

ISODATA-合并操作

① 计算当前所有类别聚类中心两两之间的距离，用矩阵D表示

② 将$D(i,j)<d_{\min}(i\neq j)$的两个类别ω_i 和ω_j 进行合并，形成一个新的类，该类的聚类中心为：

$$\boldsymbol{m}_{new}=\frac{1}{N_i+N_j}\left(N_i\boldsymbol{m}_i+N_j\boldsymbol{m}_j\right)$$

N_i和N_j为两个类别 ω_i 和 ω_j 中的样本个数，新的聚类中心可看作是对这两个类别进行加权求和。

6.4 动态聚类

6.4.3 ISODATA算法

➢ 算法流程：

ISODATA-分裂操作

① 计算每个类别下所有样本在每个维度下的方差；

② 从每个类别的所有维度方差中挑选出最大的方差$\sigma_{\max}$；

③ **如果某个类别的**$\sigma_{\max} > Sigma$ **并且该类别中所包含的样本数量**$N_i \geq 2N_{\min}$**，则进行分裂操作，**前往步骤4。如果不满足上述条件则退出分裂操作；

④ 将满足步骤3条件的类分裂成两个子类，并令$C_0 = C_0 + 1.$

6.4.3 ISODATA算法

➢ 小结

该算法能够在聚类过程中根据各个类所包含样本的实际情况动态调整聚类中心的数目。

a) 如果某个类中样本分散程度较大（通过方差进行衡量）并且样本数量较大（$N_i \geq 2N_{\min}$），则对该类进行分裂操作；

b) 如果某两个类别靠得比较近（通过聚类中心的距离衡量，$D(i,j)<d_{\min}(i\neq j)$），则对它们进行合并操作。

6.4 动态聚类

6.4.4 小结

- K-均值聚类

  基本思想：设定初始聚类中心、迭代更新、直至收敛

  问题分析：初始聚类中心选择问题、选定聚类数量问题
  （K-means++）

- ISODATA聚类

  基本思想：在K-均值聚类基础上增加对聚类结果的“合并”和“分裂”，可有效解决聚类数量需要设定的问题。