机器学习导论-无监督学习

Introduction to Machine Learning-Unsupervised Learning

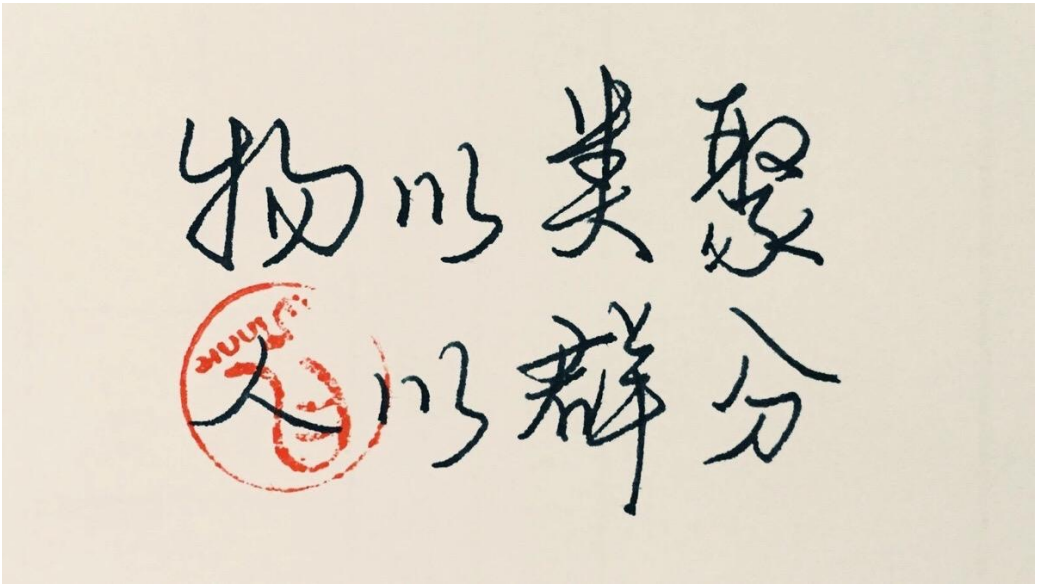
李文斌

南京大学 智能科学与技术学院

www.liwenbin.cn，liwenbin@nju.edu.cn

2025年04月14日

引入：聚类算法



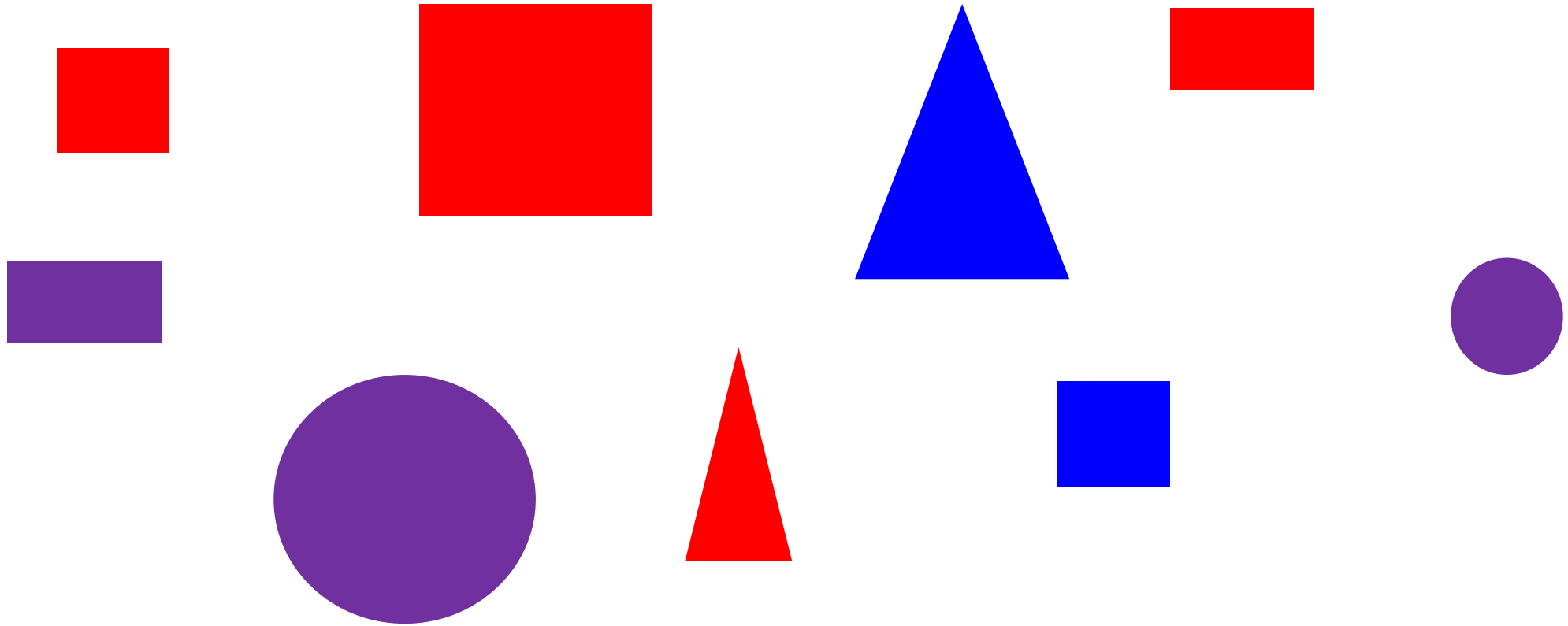
大纲

- 相关概念
- 距离度量
- 聚类准则
- 聚类方法
- 聚类评价
- 前沿进展

引入：聚类算法

南京大学 NANJING UNIVERSITY

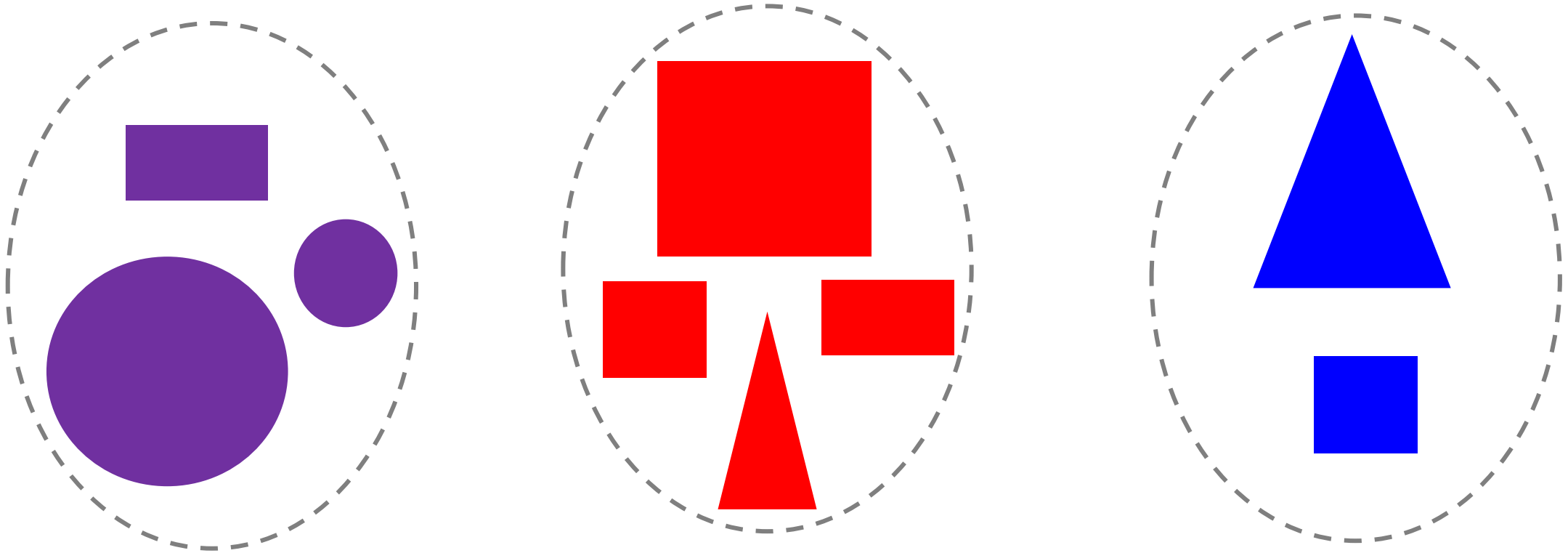
- 对下面图形进行划分（幼儿园作业）



引入：聚类算法

南京大学 NANJING UNIVERSITY

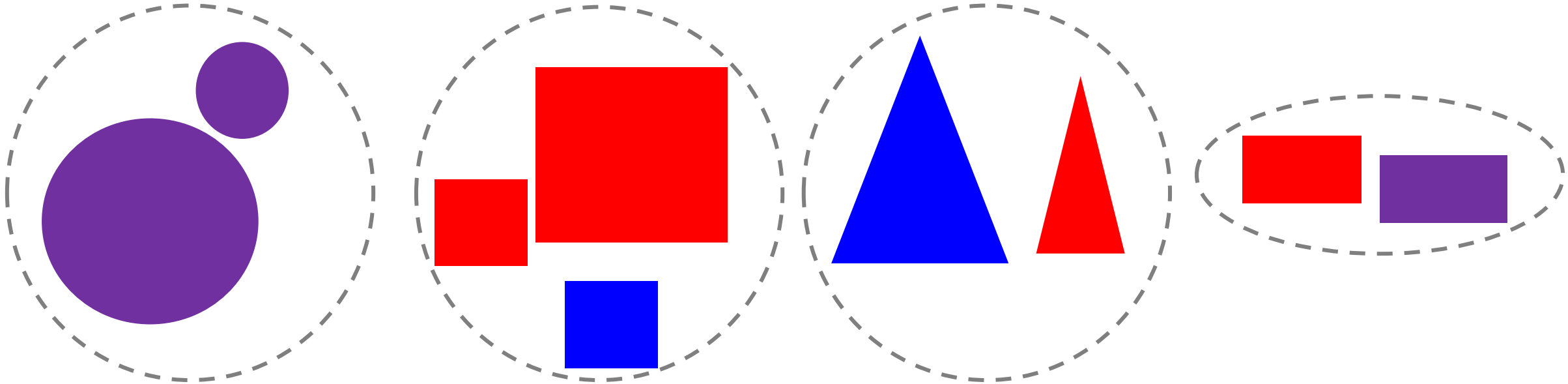
- 颜色



引入：聚类算法

南京大学 NANJING UNIVERSITY

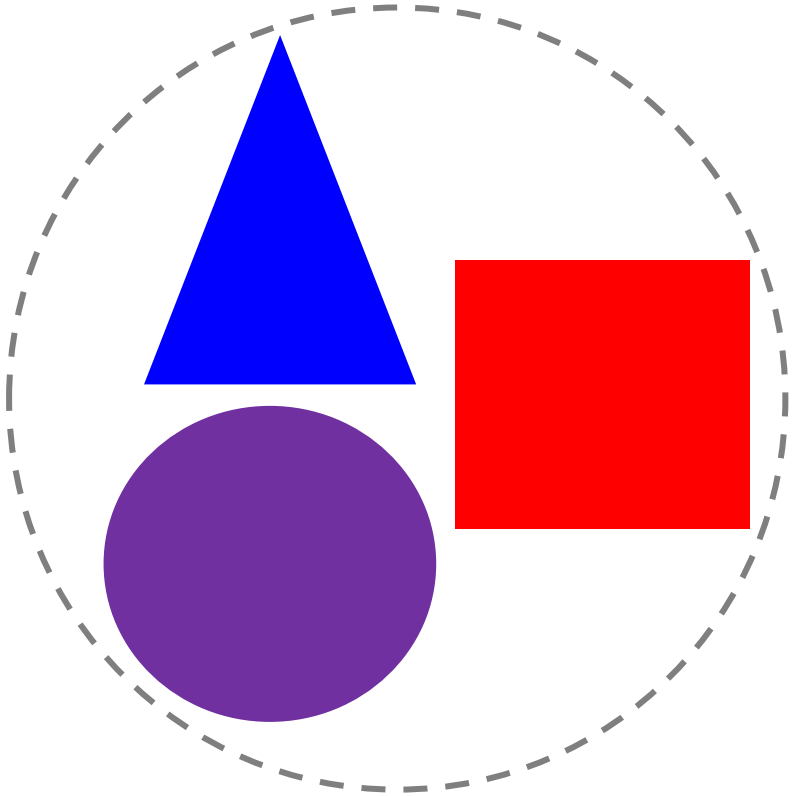
- 形状

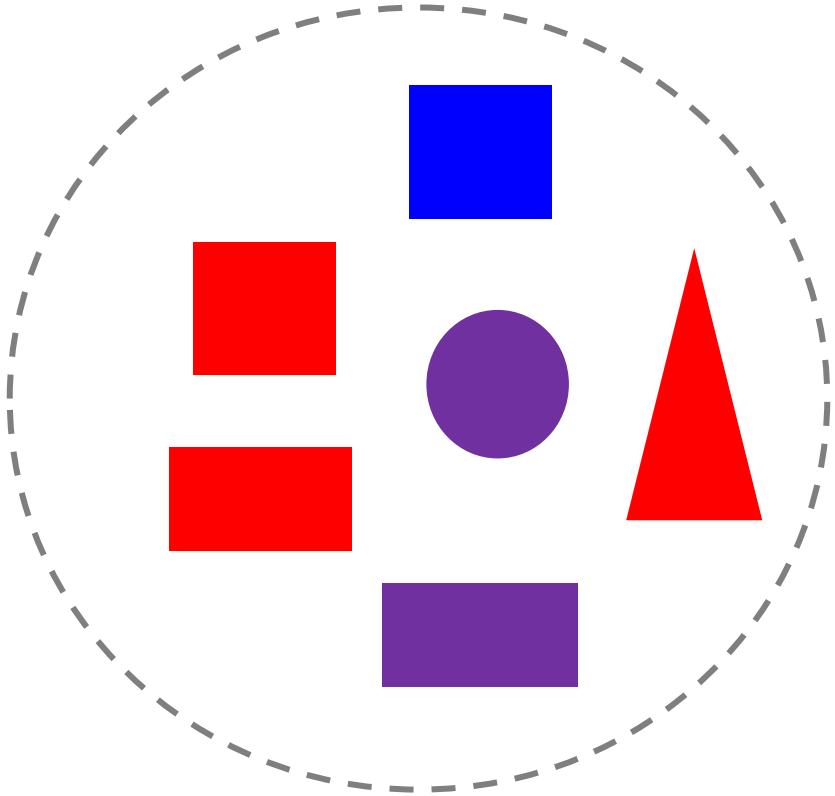


引入：聚类算法

南京大学 NANJING UNIVERSITY

- 大小

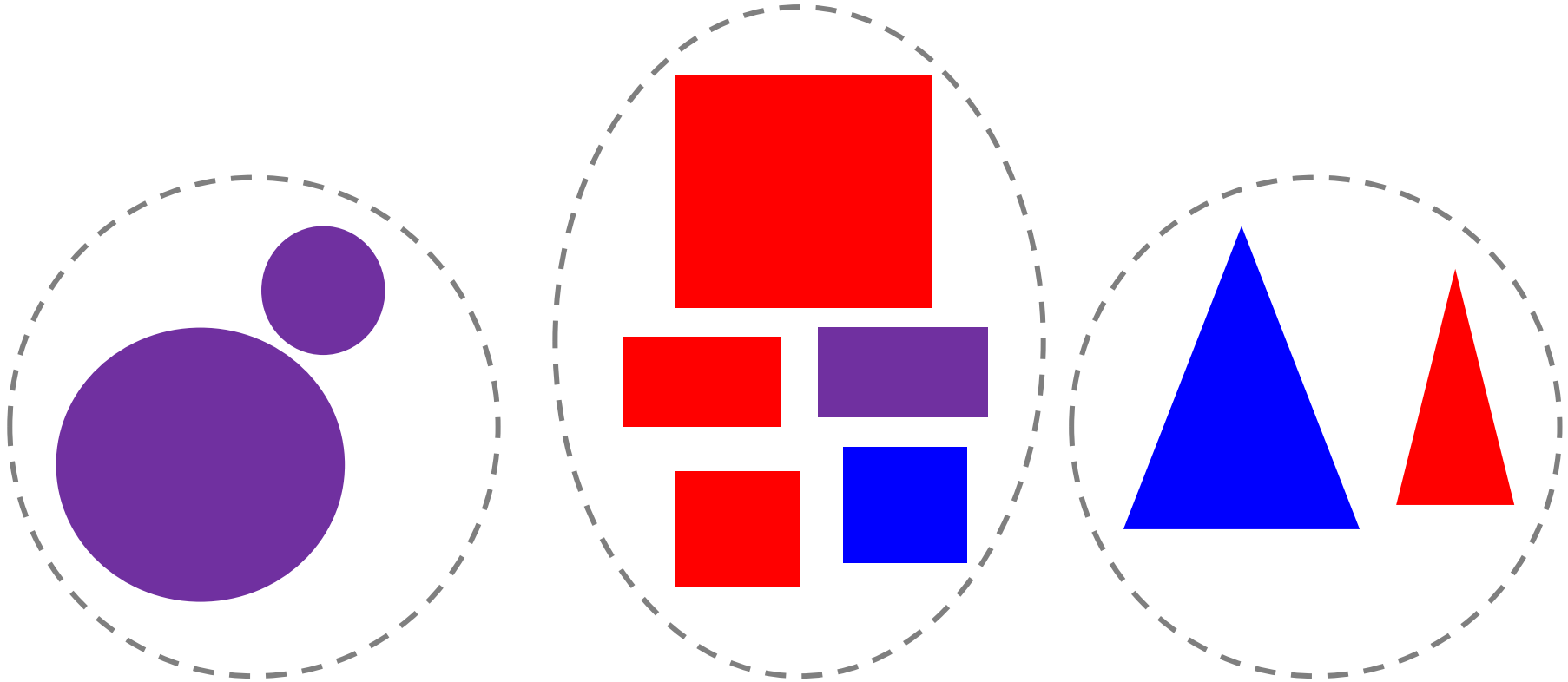




引入：聚类算法

南京大学 NANJING UNIVERSITY

- 顶点数



引入：聚类算法

南京大学 NANJING UNIVERSITY

聚类的“好坏”不存在绝对标准！

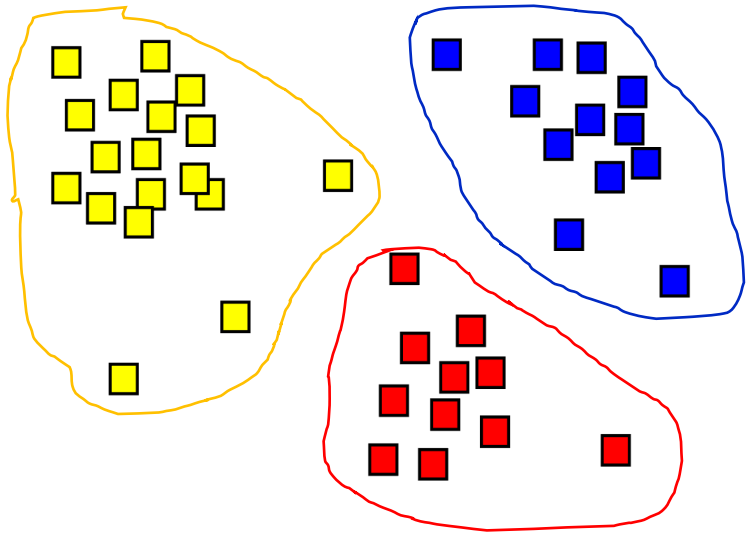
The goodness of clustering depends on the opinion of the user

聚类也许是机器学习中“新算法”出现最多、最快的领域，
总能找到一个新的“标准”，使以往算法对它无能为力

相关概念

南京大学 NANJING UNIVERSITY

- **聚类（簇、类）：数据对象的集合**
  - 在同一个簇/类中，数据对象是相似的
  - 不同簇/类之间的数据对象是不相似的



相关概念

南京大学 NANJING UNIVERSITY

❑ 聚类算法

- 根据给定的**相似性评价标准**，将一个数据集合分组/划分成几个聚类（簇）

❑ 数学形式化

- 样本集合：$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m\},\ \boldsymbol{x}_i \in \mathbb{R}^d$
- 聚类成k个簇
  - $\{C_l | l = 1,2, \dots, k\}$
  - $C_{l'} \bigcap_{l' \neq l} C_l = \emptyset$
  - $D = \bigcup_{l=1}^{k} C_l$

相关概念

南京大学 NANJING UNIVERSITY

❑ 聚类的依据

- 将整个数据集中每个样本的特征向量看成是分布在特征空间中的一个点，点与点之间的距离即可作为相似性度量依据

- 聚类分析是根据不同样本之间的差异，根据距离函数的规律（大小）进行聚类的

相关概念

南京大学 NANJING UNIVERSITY

❑ 一个好的聚类算法

- 聚类（簇）内部高相似性
- 聚类（簇）之间低相似性

相关概念

南京大学 NANJING UNIVERSITY

❑ 聚类算法中“类”的特征：

- 聚类所说的类不是事先给定的，而是根据数据的相似性和距离来划分

  （**无监督的算法**）

- 聚类的数目和结构都没有事先假定

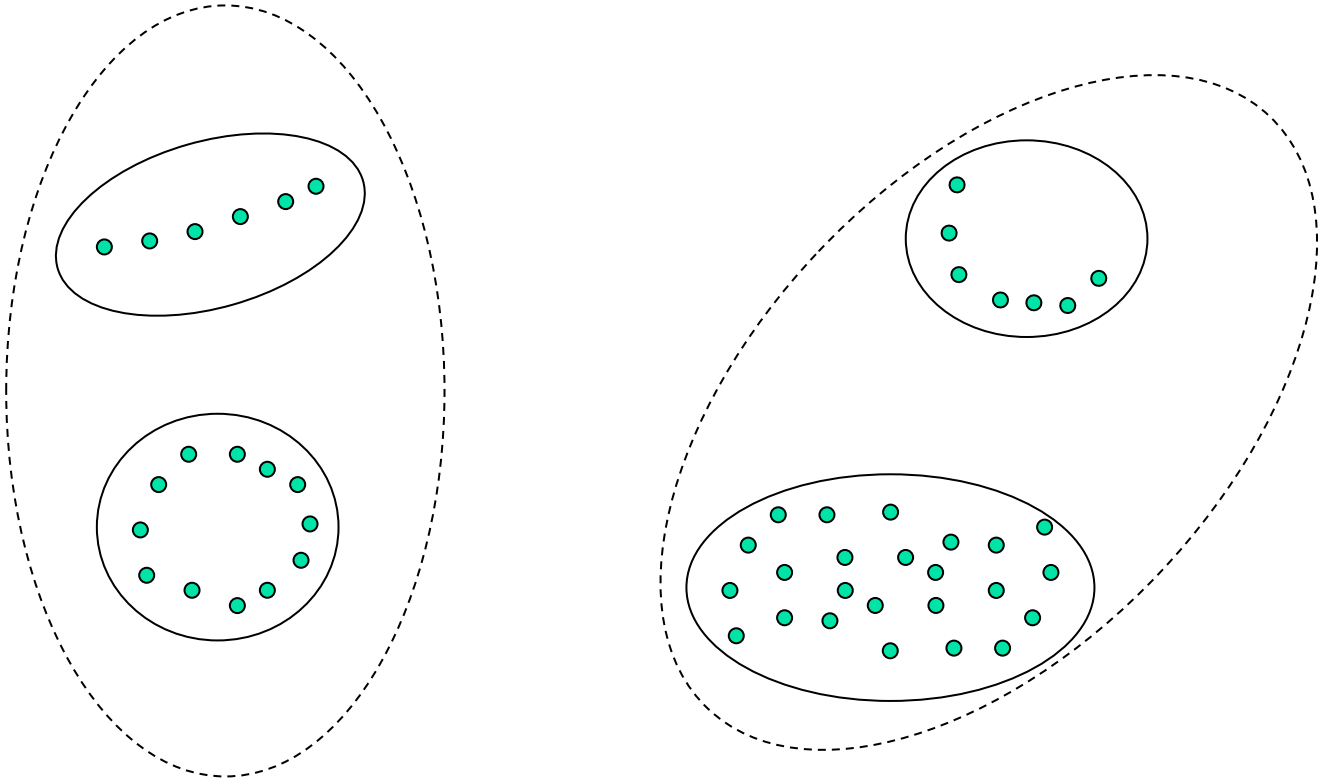
❑ 聚类方法的目的是寻找数据中：

- 潜在的**自然分组结构**
- 感兴趣的**关系**

相关概念

南京大学 NANJING UNIVERSITY

- 粒度对聚类的影响



数据的粗聚类是2类,细聚类为4类

相关概念

南京大学 NANJING UNIVERSITY

❑ 聚类的关键

- 特征的选取或设计
- 距离度量函数的选择

❑ 聚类分析的有效性：

- 若数据点的分布是一群一群的，同一群样本密集（距离很近），不同群样本距离很远，则很容易聚类；
- 若样本集的分布聚成一团，不同群的样本混在一起，则很难分类；
- 对具体数据做聚类分析的**关键是选取合适的特征**。特征选取得好，容易区分，选取得不好，很难区分。

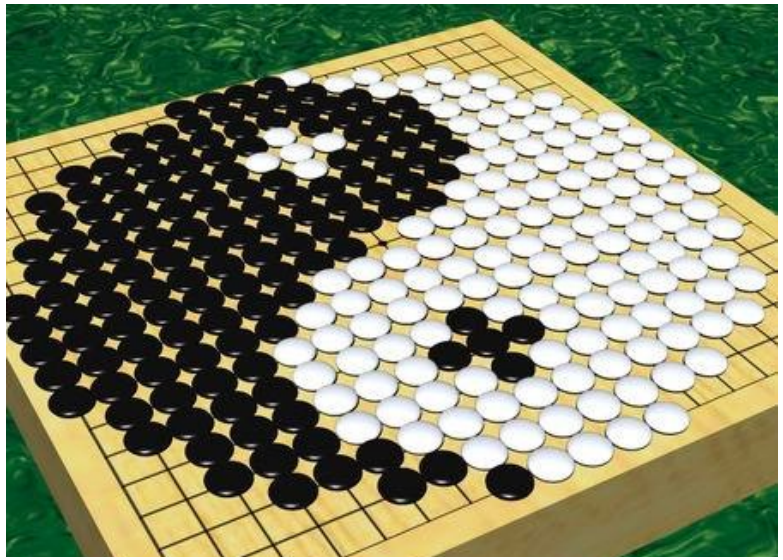
聚类分析方法是否有效，与数据分布形式有很大关系！

相关概念

南京大学 NANJING UNIVERSITY

❑ 两类聚类实例：一摊黑白围棋子

- 选**颜色**作为特征进行聚类，用“1”代表白，“0”代表黑，则很容易分类；
- 选**大小**作为特征进行聚类，则白子和黑子的特征相同，不能分类（把白子和黑子分开）



大纲

- 相关概念
- 距离度量
- 聚类准则
- 聚类方法
- 聚类评价
- 前沿进展

距离度量

南京大学 NANJING UNIVERSITY

- **目的：度量同类样本间的相似性和不同类样本间的差异性**
- **度量函数和度量空间**
  - 度量空间是一个有序对，记作(X,d), 其中X是一个集合，d是 X 上的度量函数：它把X 中的每一对点$\boldsymbol{x}$, $\boldsymbol{y}$映射到一个非负实数，并满足以下四条公理：
  - 非负性：$d(\boldsymbol{x},\boldsymbol{y}) \geq 0$
  - 唯一性：$d(\boldsymbol{x},\boldsymbol{y}) = 0 \leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
  - 对称性：$d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x})$
  - 三角不等式：$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \boldsymbol{X}, d(\boldsymbol{x},\boldsymbol{z}) \leq d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z})$

距离度量

南京大学 NANJING UNIVERSITY

❑ 常用度量函数（$\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^n$）

- 欧氏距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j)}$
- 余弦相似性：$\mathrm{s}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i{}^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_\mathrm{j}\|}$
- 曼哈顿距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$
- 切比雪夫距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty$
- 马氏距离：$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$

距离度量

南京大学 NANJING UNIVERSITY

❑ 常用度量函数（$\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^n$）

- 欧式距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$
- 曼哈顿距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$
- 切比雪夫距离：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty$

闵可夫斯基距离
Minkowski distance

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

聚类准则

南京大学 NANJING UNIVERSITY

❑ 类的定义

- 类的定义有很多种，类的划分具有人为规定性。一个聚类结果的优劣最后只能根据实际来评价。

- 定义之一：设集合S中任意元素$\boldsymbol{x}_i$与$\boldsymbol{x}_j$间的距离有:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq h$$

其中h为给定的阀值，称S对于阀值h组成一类。

聚类准则

南京大学 NANJING UNIVERSITY

❑ 类的定义

- 聚类有了样本的相似性度量/距离度量，还需要一种基于数值的聚类准则，才能将相似的样本分在同一类，相异的样本分在不同的类
- 判别聚类结果好坏的一般标准：类内距离小，类间距离大；或者，

  簇内相似度（intra-cluster similarity）高，

  簇间相似低（inter-cluster similarity）低。
- 需要一个能对聚类过程或聚类结果的优劣进行评估的准则函数。如果聚类准则函数选择得好，聚类质量就会高。

大纲

- 相关概念
- 距离度量
- **聚类准则**
- 聚类方法
- 聚类评价
- 前沿进展

聚类准则

南京大学 NANJING UNIVERSITY

❑ 试探方法

- 凭直观感觉或经验，针对实际问题定义一种距离度量的阈值，然后按最近邻规则指定某些样本属于某一个聚类类别。
- 例如对欧氏距离，它反映了样本间的近邻性，但将一个样本分到不同类别中的哪一个时，还必须规定一个距离度量的阈值作为聚类的判别准则。

聚类准则

南京大学 NANJING UNIVERSITY

❑ 聚类准则函数方法

- 依据：由于聚类是将样本进行分类以使类别之间的分离性尽可能大，因此聚类准则应是反映类别间相似性或分离性的函数；
- 每个类别都是由一系列样本组成的，因此一般来说类别（sample sets）的可分离性和样本（samples）的可分离性是直接相关的；
- 可以定义聚类准则函数为样本集$\{\boldsymbol{x}\}$和类别$\{S_j, j = 1,2,\cdots,c\}$的函数，从而使聚类分析转化为寻找准则函数极值的最优化问题。

聚类准则

南京大学 NANJING UNIVERSITY

❑ 聚类准则函数方法

- **一种聚类准则函数 J 的定义**

$$J = \sum_{\mathrm{j}=1}^{\mathrm{c}} \sum_{\mathrm{x} \in \mathrm{S_j}} \|\mathrm{x} - \mathrm{m_j}\|^2$$

- J 代表了属于 c 个聚类类别全部样本与其相应类别均值m_j之间的误差平方和
- 对于不同的聚类形式，J 值是不同的
- 目的：求取使 J 值达到最小的聚类形式

大纲

- 相关概念
- 距离度量
- 聚类准则
- 聚类方法
- 聚类评价
- 前沿进展

聚类方法

- 试探聚类
- 原型聚类
- 密度聚类
- 层次聚类

聚类方法

南京大学 NANJING UNIVERSITY

- **试探聚类** ——→ **早期聚类方法**
- 原型聚类
- 密度聚类
- 层次聚类

基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 按最近邻规则的简单试探法

- 给定 N 个待分类的数据样本$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$，要求按距离阈值 T，将它们分类到聚类中心 $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$

- **第一步**：
  - 任取一样本$\boldsymbol{x}_i$作为一个聚类中心的初始值，例如令$\boldsymbol{z}_1 = \boldsymbol{x}_1$
  - 计算 D_{21}=‖ $\boldsymbol{x}_2 - \boldsymbol{z}_1$ ‖$_2$
  - 若$D_{21} > T$，则确定一个新的聚类中心 $\boldsymbol{z}_2 = \boldsymbol{x}_2$
  - 否则$\boldsymbol{x}_2$属于以$\boldsymbol{z}_1$为中心的聚类

- **第二步**：
  - 假设已经有聚类中心$\boldsymbol{z}_1, \boldsymbol{z}_2$
  - 计算D_{31} =‖ $\boldsymbol{x}_3 - \boldsymbol{z}_1$ ‖$_2$， D_{32} =‖ $\boldsymbol{x}_3 - \boldsymbol{z}_2$ ‖$_2$
  - 若$D_{31} > T$且$D_{32} > T$，则得一个新的聚类中心 $\boldsymbol{z}_3 = \boldsymbol{x}_3$
  - 否则$\boldsymbol{x}_3$属于离$\boldsymbol{z}_1$和$\boldsymbol{z}_2$中的最近者
  - ……
  - 如此重复下去，直至将N个样本分类完毕。

基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 按最近邻规则的简单试探法

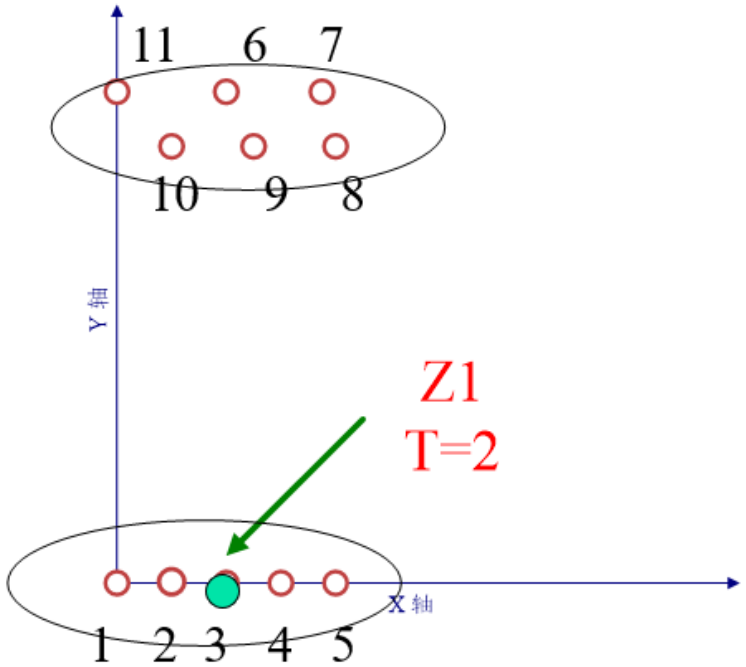
- 在实际中，对于高维样本很难获得准确的先验知识，因此只能选用不同的阈值和起始点来试探，所以这种方法在很大程度上依赖于以下因素：
  - 第一个聚类中心的位置
  - 待聚类样本的排列次序
  - 距离阈值 T 的大小
  - 样本分布的几何性质

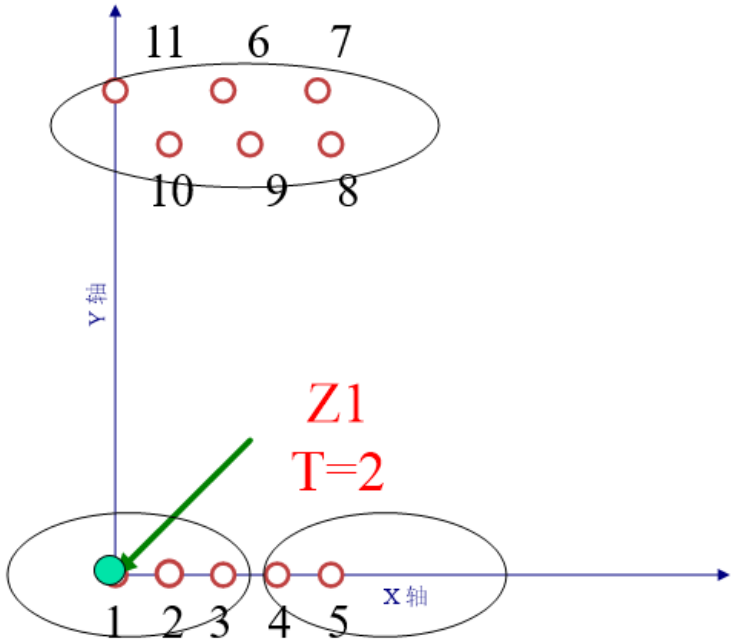
基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 按最近邻规则的简单试探法

- **初始点不同**



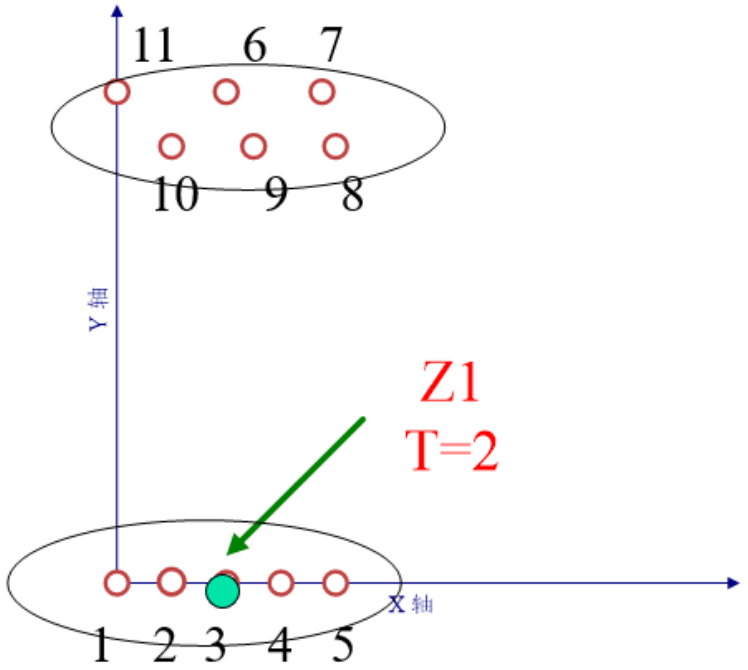


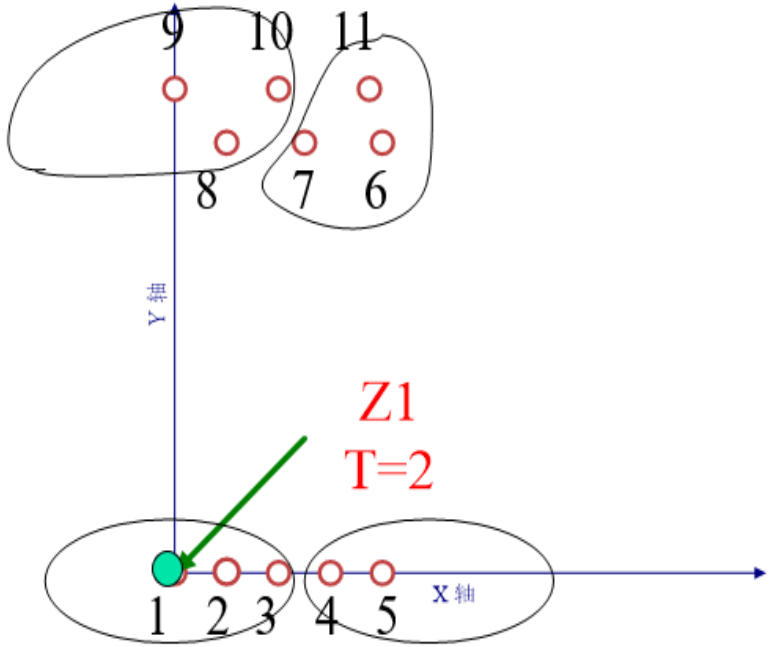
基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 按最近邻规则的简单试探法

- **样本次序不同**



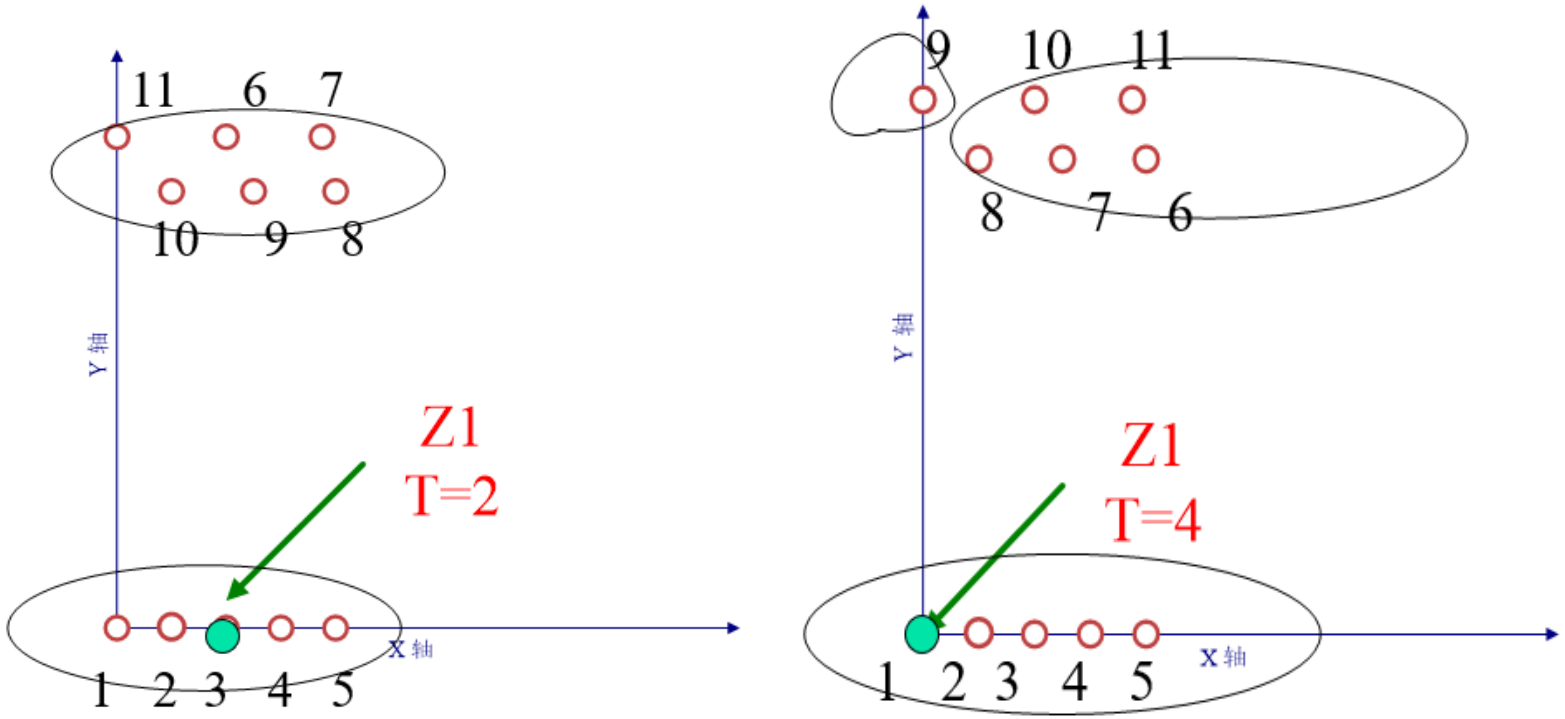


基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 按最近邻规则的简单试探法

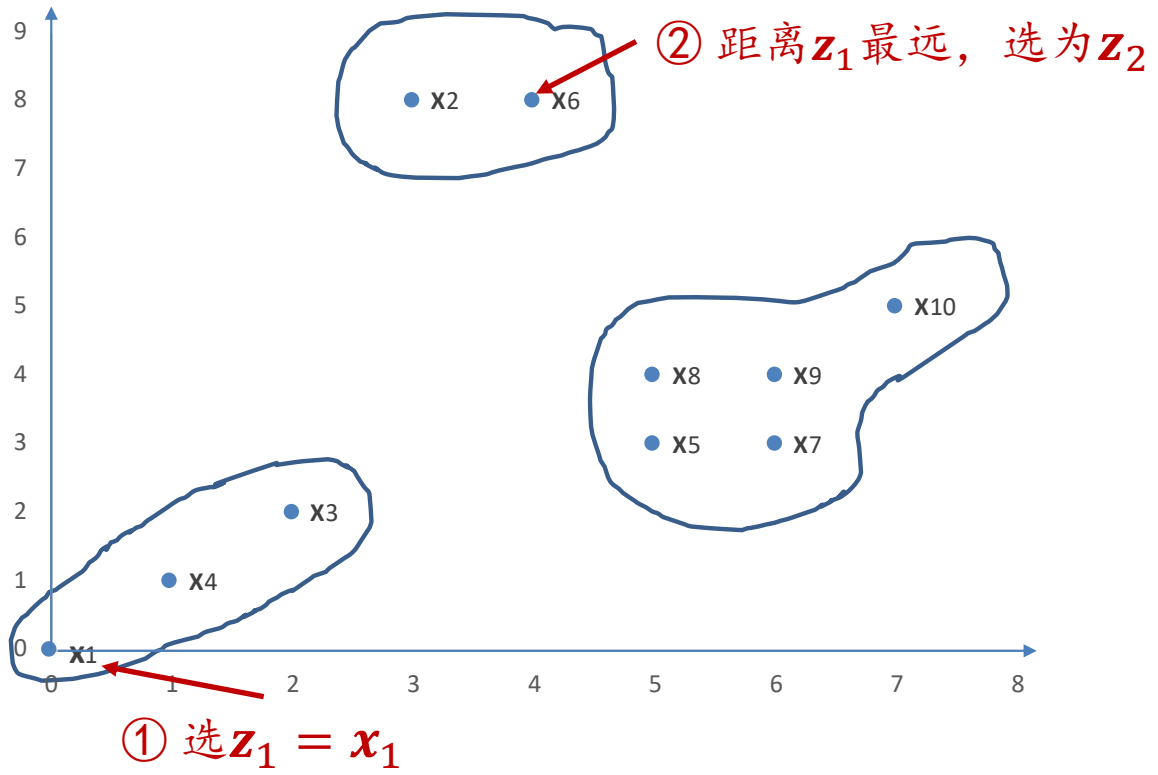
- 阈值 T 不同



基于试探的聚类搜索算法

❑ 最大最小距离思想

- **基本思想**：以试探类间欧氏距离为最大作为预选出聚类中心的条件



基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 第一步

选任意一个样本作为第一个聚类中心，如$\boldsymbol{z}_1 = \boldsymbol{x}_1$

❑ 第二步

选距离$\boldsymbol{z}_1$最远的样本作为第二个聚类中心。经计算，$\|\boldsymbol{x}_6 - \boldsymbol{z}_1\|$最大，所以$\boldsymbol{z}_2 = \boldsymbol{x}_6$

❑ 第三步

逐个计算各样本$\{\boldsymbol{x}_i, i = 1,2,\cdots,N\}$与$\{\boldsymbol{z}_1, \boldsymbol{z}_2\}$之间的距离，即 $D_{i1} = \|\boldsymbol{x}_i - \boldsymbol{z}_1\|$, $D_{i2} = \|\boldsymbol{x}_i - \boldsymbol{z}_2\|$

并选出其中的最小距离$\min(D_{i1}, D_{i2}), i = 1,2,\cdots,N$

❑ 第四步

在所有样本的最小值中选出最大距离，若该最大值达到$\|\boldsymbol{z}_2 - \boldsymbol{z}_1\|$的一定比例以上，则相应的样本点取为第三个聚类中心$\boldsymbol{z}_3$，即

若$\max\{\min(D_{i1}, D_{i2}), i = 1,2,\cdots,N\} > \theta\|\boldsymbol{z}_2 - \boldsymbol{z}_1\|$，则$\boldsymbol{z}_3 = \boldsymbol{x}_i$

否则，若找不到适合要求的样本作为新的聚类中心，则找聚类中心的过程结束。

这里，θ 可用试探法取一固定分数，如1/2。

基于试探的聚类搜索算法

南京大学 NANJING UNIVERSITY

❑ 第五步

若有$\boldsymbol{z}_3$存在，则计算$\max\{\min(D_{i1}, D_{i2}, D_{i3}), i = 1,2,\cdots,N\}$。若该值超过$\|\boldsymbol{z}_2 - \boldsymbol{z}_1\|$的一定比例，则存在$\boldsymbol{z}_4$，否则找聚类中心的过程结束。

在此例中，无$\boldsymbol{z}_4$满足条件。

❑ 第六步

将样本$\{\boldsymbol{x}_i, i = 1,2,\cdots,N\}$按最近距离分到最近的聚类中心：

- $\boldsymbol{z}_1 = \boldsymbol{x}_1 : \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4\}$为第一类
- $\boldsymbol{z}_2 = \boldsymbol{x}_6 : \{\boldsymbol{x}_2, \boldsymbol{x}_6\}$为第二类
- $\boldsymbol{z}_3 = \boldsymbol{x}_7 : \{\boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}\}$为第三类

最后，还可在每一类中计算各样本的均值，得到更具代表性的聚类中心。

早期聚类方法缺点

- 高度依赖于初始值选择
- 无法有效处理高维数据，对噪声和离群点极为敏感
- 缺乏良好的优化目标函数，缺乏理论保证

聚类方法

南京大学 NANJING UNIVERSITY

- 试探聚类
- **原型聚类**
- **密度聚类**
- **层次聚类**

原型聚类法

南京大学 NANJING UNIVERSITY

- **基于原型的聚类**（Prototype-based clustering）
- **假设**：聚类结构能通过一组原型刻画
- **过程**：先对原型初始化，然后对原型进行迭代更新求解
- **代表**：k-均值聚类，ISODATA，高斯混合聚类

K-means算法

南京大学 NANJING UNIVERSITY

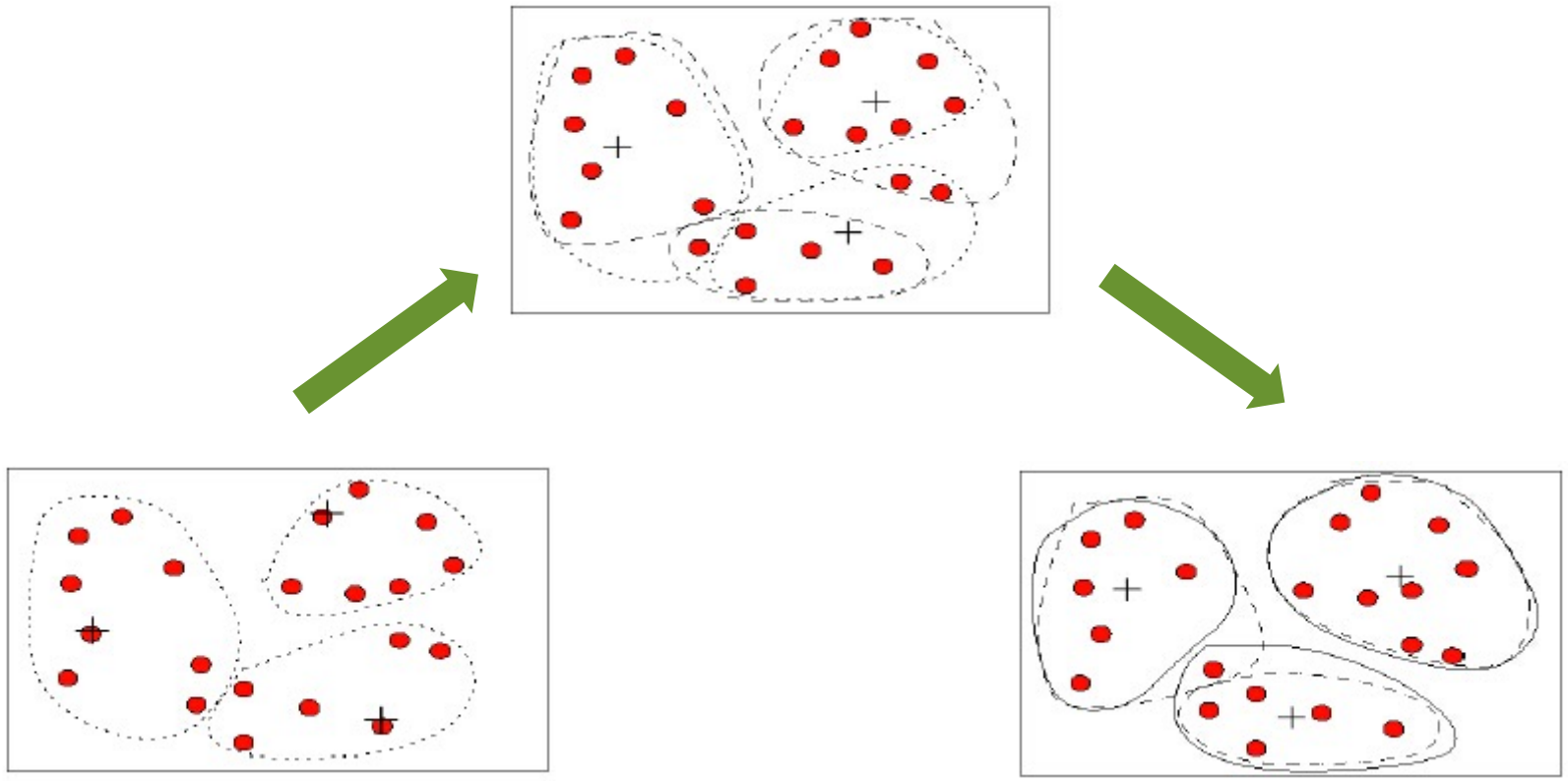
❑ 算法流程

- **Step1**：选择一个聚类数量 k
- **Step2**：初始化聚类中心 $\mu_1,\ldots,\mu_k$

  随机选择k个样本点，设置这些样本点为中心
- **Step3**：对每个样本点, 计算样本点到 k 个聚类中心的距离（使用某种距离度量），

  将样本点分距离它最近的聚类中心所属的类簇
- **Step4**：重新计算聚类中心，聚类中心为属于这一个类簇的所有样本的均值
- **Step5**：如果没发生样本所属的类簇改变的情况，则退出，否则，返回Step3继续

K-means算法



K-means算法

南京大学 NANJING UNIVERSITY

❑ 讨论

- *K*-means算法的结果受如下选择的影响：
  - 所选聚类的数目
  - 聚类中心的初始分布
  - 样本分布的几何性质
  - ……
- 在实际应用中，需要试探不同的*K*值和选择不同的聚类中心的起始值
- 如果数据样本可以形成若干个相距较远的孤立的区域分布，一般都能得到较好的收敛效果
- *K*-means算法比较适合于分类数目已知的情况

K-means++算法

南京大学 NANJING UNIVERSITY

- **基本思想：K 个初始聚类中心相互之间应该分得越开越好**

K-means++算法
Step 1:从数据集中随机选取一个样本作为初始聚类中心$\boldsymbol{c}_1$;
Step 2:首先计算每个样本$\boldsymbol{x}$与当前已有聚类中心之间的最短距离（即与最近的一个聚类中心的距离），用$D(\boldsymbol{x})$表示；接着计算每个样本被选为下一个聚类中心的概率$\frac{D(\boldsymbol{x})^2}{\sum_{x\in X} D(\boldsymbol{x})^2}$。最后，按照轮盘法选择出下一个聚类中心；
Step 3:重复第2步直到选择出共K个聚类中心；
之后的过程与经典K-means算法中第2步至第4步相同

ISODATA算法

南京大学 NANJING UNIVERSITY

❑ 迭代自组织数据分析算法

- Iterative Self-organizing Data Analysis Techniques

❑ 基本步骤和思路

- （1）选择某些初始值。可选不同的参数，也可在迭代过程中人为修改，以将N个样本按指标分配到各个聚类中心中去。
- （2）计算各类中诸样本的距离指标函数。
- （3）~（6）按给定的要求，将前一次获得的聚类集合进行分裂和合并处理，（5）为分裂处理，（6）为合并处理，从而获得新的聚类中心。
- （7）重新进行迭代运算，计算各项指标，判断聚类结果是否符合要求。经过多次迭代后，若结果收敛，则运算结束。

ISODATA算法

南京大学 NANJING UNIVERSITY

❑ 过程中能根据各个类别的实际情况进行**分裂**和**合并**两种操作来调整聚类中心数

ISODATA算法
Step 1:从数据集中随机选取K_0个样本作为初始聚类中心$C=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_{K_0}\}$;
Step 2:针对数据集中每个样本$\boldsymbol{x}_i$,计算它到K_0个聚类中心的距离并将其分到距离最小的聚类中心所对应的类中;
Step 3:判断上述每个类中的元素数目是否小于N_{min}。如果小于N_{min}则需要丢弃该类，令$K=K-1$，并将该类中的样本重新分配给剩下类中距离最小的类;
Step 4:针对每个类别$\boldsymbol{c}_i$,重新计算它的聚类中心$\boldsymbol{c}_i=\frac{1}{\lvert c_i\rvert}\sum_{\boldsymbol{x}\in c_i}\boldsymbol{x}$（即属于该类的所有样本的质心）;
Step 5:如果当前$K\leq\frac{K_0}{2}$，说明当前类别数太少，前往**分裂操作**;
Step 6:如果当前$K\geq 2K_0$，说明当前类别数太多，前往**合并操作**;
Step 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;

ISODATA算法

南京大学 NANJING UNIVERSITY

ISODATA-合并操作
Step 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，其中$D(i,i)=0$;
Step 2:对于$D(i,j)<d_{min}(i\neq j)$的两个类别需要进行**合并操作**，变成一个新的类，该类的聚类中心位置为： $$\boldsymbol{m}_{new}=\frac{1}{n_i+n_j}\left(n_i\boldsymbol{m}_i+n_j\boldsymbol{m}_j\right)$$ 上式中的n_i和n_j表示这两个类别中的样本个数，$\boldsymbol{m}_i$和$\boldsymbol{m}_j$分别表示第i和j个类别中心，新的聚类中心可以看作是对这两个类别进行加权求和。

两个类别对应聚类中心之间所允许最小距 d_{min}：是否进行合并的阈值

ISODATA算法

ISODATA-分裂操作
Step 1:计算每个类别下所有样本在每个属性维度下的方差；
Step 2:针对每个类别的所有方差挑选出最大的方差σ_{max}；
Step 3:如果某个类别的$\sigma_{max} > Sigma$，并且该类别所包含的样本数量$n_i \geq 2n_{min}$,则可以进行分裂操作，前往步骤4。如果不满足上述条件则退出分裂操作；
Step 4:将满足步骤3中条件的类分裂成两个子类别并令$K = K + 1$，h是一个比例系数，在方差最大的属性维度（假设为第k维）上正负偏移$h \cdot \sigma_{max}$ $$m_{ik}^{(+)} = m_{ik} + h \cdot \sigma_{max}, m_{ik}^{(-)} = m_{ik} - h \cdot \sigma_{max}.$$

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

ISODATA算法

南京大学 NANJING UNIVERSITY

❑ 与*K*-means算法比较

- *K*-means算法通常适合于类别数目已知的聚类，而ISODATA算法则更加灵活；
- 从算法角度看，ISODATA算法与*K*-means算法相似，聚类中心都是通过样本均值的迭代运算来决定的；
- ISODATA算法加入了一些试探步骤，并且可以结合人机交互的结构，使其能利用中间结果所取得的经验更好地进行分类；
- ISODATA原理非常直观，不过它需要额外指定较多的参数，并且某些参数同样很难准确指定出一个较合理的值，因此ISODATA算法在实际过程中并没有特别受欢迎。

高斯混合聚类

❑ 高斯混合聚类 (Gaussian Mixture Clustering)

采用概率模型来表达聚类原型

n 维样本空间中的随机向量 $\boldsymbol{x}$ 若服从高斯分布，则其概率密度函数为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

假设样本由下面这个高斯混合分布生成：

$$p_{\mathcal{M}}(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x} \mid \mu_i, \boldsymbol{\Sigma}_i)$$

- 根据 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 定义的先验分布选择高斯混合成分，其中 α_i 为选择第 i 个混合成分的概率；
- 然后，根据被选择的混合成分的概率密度函数进行采样，从而生成相应的样本

高斯混合聚类

南京大学 NANJING UNIVERSITY

- 样本$\boldsymbol{x}_j$由第i个高斯混合成分生成的后验概率为：

$$p_{\mathcal{M}}(z_j = i|\boldsymbol{x}_j) = \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\boldsymbol{x}_j|z_j = i)}{p_{\mathcal{M}}(\boldsymbol{x}_j)} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j|\mu_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j|\mu_l, \boldsymbol{\Sigma}_l)}$$

简记为 $\gamma_{ij}(i = 1,2,\cdots,k)$

- 参数估计可采用极大似然法，考虑最大化对数似然

$$LL(D) = \ln\left(\prod_{j=1}^{m} p_{\mathcal{M}}(\boldsymbol{x}_j)\right) = \sum_{j=1}^{m} \ln\left(\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \mu_i, \boldsymbol{\Sigma}_i)\right)$$

EM 算法：

- (E步) 根据当前参数计算每个样本属于每个高斯成分的后验概率 γ_{ji}
- (M步) 更新模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \le i \le k)\}$

密度聚类法

南京大学 NANJING UNIVERSITY

- **基于密度的聚类** (density-based clustering)
- **假设**：聚类结构能通过样本分布的紧密程度确定
- **过程**：从样本密度的角度来考察样本之间的可连接性，

  并基于可连接样本不断扩展聚类簇
- **代表**：DBSCAN, OPTICS, DENCLUE

DBSCAN

南京大学 NANJING UNIVERSITY

❑ 关键概念

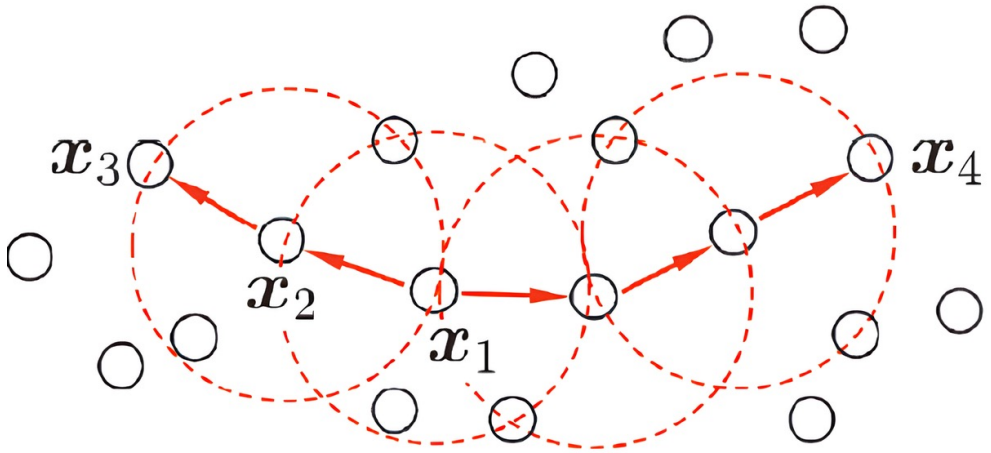
- **核心对象**(core object): 若$\boldsymbol{x}_j$的 ϵ-邻域至少包含$MinPts$个样本，即$\left|N_\epsilon(\boldsymbol{x}_j)\right| \geq MinPts$，则$\boldsymbol{x}_j$是一个核心对象；
- **密度直达**(directly density-reachable): 若$\boldsymbol{x}_j$位于$\boldsymbol{x}_i$的 ϵ-邻域中，且$\boldsymbol{x}_i$是核心对象，则称$\boldsymbol{x}_j$由$\boldsymbol{x}_i$密度直达；
- **密度可达**(density-reachable): 对$\boldsymbol{x}_i$与$\boldsymbol{x}_j$，若存在样本序列$\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n$，其中 $\boldsymbol{p}_1 = \boldsymbol{x}_i$，$\boldsymbol{p}_n = \boldsymbol{x}_j$且 $\boldsymbol{p}_{i+1}$由 $\boldsymbol{p}_i$密度直达，则$\boldsymbol{x}_j$由$\boldsymbol{x}_i$密度可达；
- **密度相连**(density-connected): 对$\boldsymbol{x}_i$与$\boldsymbol{x}_j$，若存在$\boldsymbol{x}_k$使得$\boldsymbol{x}_i$与$\boldsymbol{x}_j$均由$\boldsymbol{x}_k$密度可达，则称$\boldsymbol{x}_i$与 $\boldsymbol{x}_j$密度相连。

DBSCAN

❑ 一个例子

令$MinPts = 3$，虚线显示出ϵ-邻域，$\boldsymbol{x}_1$是一个核心对象

- $\boldsymbol{x}_2$由$\boldsymbol{x}_1$密度直达
- $\boldsymbol{x}_3$由$\boldsymbol{x}_1$密度可达
- $\boldsymbol{x}_3$由$\boldsymbol{x}_4$密度相连



DBSCAN

南京大学 NANJING UNIVERSITY

❑ 算法流程

ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

Input: *DB*: Database
Input: ε: Radius
Input: *minPts*: Density threshold
Input: *dist*: Distance function
Data: *label*: Point labels, initially *undefined*

```
1  foreach point p in database DB do                          // Iterate over every point
2      if label(p) ≠ undefined then continue                  // Skip processed points
3      Neighbors N ← RangeQuery(DB, dist, p, ε)               // Find initial neighbors
4      if |N| < minPts then                                   // Non-core points are noise
5          label(p) ← Noise
6          continue
7      c ← next cluster label                                 // Start a new cluster
8      label(p) ← c
9      Seed set S ← N \ {p}                                   // Expand neighborhood
10     foreach q in S do
11         if label(q) = Noise then label(q) ← c
12         if label(q) ≠ undefined then continue
13         Neighbors N ← RangeQuery(DB, dist, q, ε)
14         label(q) ← c
15         if |N| < minPts then continue                      // Core-point check
16         S ← S ∪ N
```

➢ **阅读参考**：https://dl.acm.org/doi/pdf/10.1145/3068335

DBSCAN

南京大学 NANJING UNIVERSITY

❑ 算法流程

ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

Input: *DB*: Database
Input: ε: Radius
Input: *minPts*: Density threshold
Input: *dist*: Distance function
Data: *label*: Point labels, initially *undefined*

ALGORITHM 2: Abstract DBSCAN Algorithm

```
1 Compute neighbors of each point and identify core points        // Identify core points
2 Join neighboring core points into clusters                      // Assign core points
3 foreach non-core point do
4 |   Add to a neighboring core point if possible                 // Assign border points
5 |   Otherwise, add to noise                                     // Assign noise points
```

```
12 |   |   if label(q) ≠ undefined then continue
13 |   |   Neighbors N ← RangeQuery(DB, dist, q, ε)
14 |   |   label(q) ← c
15 |   |   if |N| < minPts then continue                        // Core-point check
16 |   |   S ← S ∪ N
```

➢ **阅读参考**：https://dl.acm.org/doi/pdf/10.1145/3068335

层次聚类法

南京大学 NANJING UNIVERSITY

❑ 基本思想

- 将数据样本按距离准则逐步分类，类别由多到少，直到获得合适的分类要求为止

❑ 距离准则函数

- 进行聚类合并的一个关键就是每次迭代中形成的聚类之间以及它们和样本之间距离的计算，采用不同的距离函数会得到不同的计算结果

❑ 主要的距离计算准则：

- ➢ 最短距离法 （两个集合所有距离最小值）
- ➢ 最长距离法 （两个集合所有距离最大值）
- ➢ 类平均距离法（两个集合所有距离平均值）

层次聚类法

南京大学 NANJING UNIVERSITY

❑ 算法流程

- **第一步**：设初始样本共有N个，每个样本自成一类，即建立N类，$G_1^{(0)}, G_2^{(0)}, \cdots, G_N^{(0)}$。计算各类之间的距离（初始为各样本间的距离），得到一个$N * N$维的距离矩阵$D^{(0)}$。这里，标号(0)表示聚类开始运算前的状态。
- **第二步**：假设前一步聚类运算中已求得距离矩阵$D^{(n)}$，n为逐次聚类合并的次数，则求$D^{(n)}$中的最小元素。如果它是$G_i^{(n)}$和$G_j^{(n)}$两类之间的距离，则将$G_i^{(n)}$和$G_j^{(n)}$两类合并为一类$G_{ij}^{(n+1)}$，由此建立新的分类：$G_1^{(n+1)}, G_2^{(n+1)}, \cdots$。
- **第三步**：计算合并后新类别之间的距离，得$D^{(n+1)}$。计算$G_{ij}^{(n+1)}$与其它没有发生合并的$G_1^{(n+1)}, G_2^{(n+1)}, \cdots$之间的距离，可采用多种不同的距离计算准则进行计算。
- **第四步**：返回第二步，重复计算及合并，直到得到满意的分类结果。（如：达到所需的聚类数目，或$D^{(n)}$中的最小分量超过给定阈值D等。）

AGNES

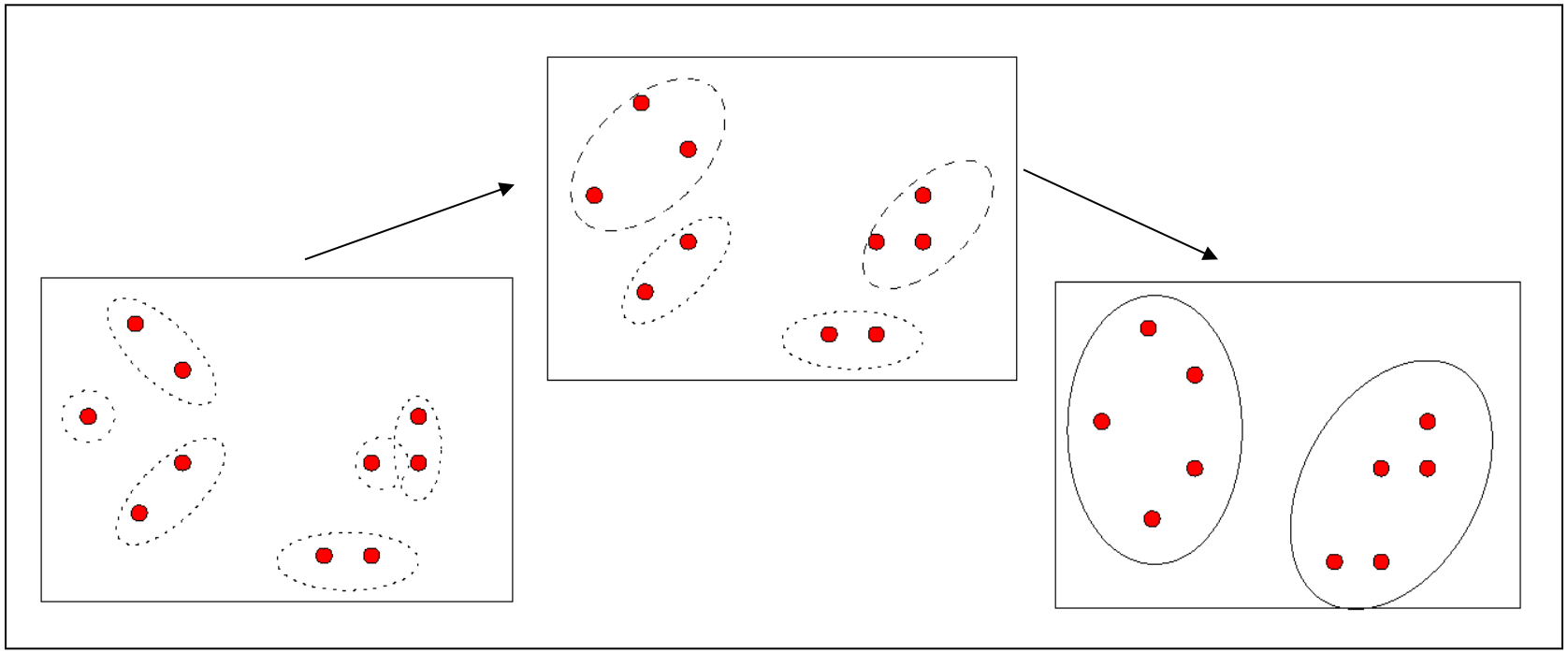
❑ AGNES (AGglomerative NESting)

在不同层次对数据集进行划分，从而形成树形的聚类结构

Step1: 将每个样本点作为一个簇

Step2: 合并最近的两个簇

Step3: 若所有样本点都存在于一个簇中，则停止；否则转到 Step2



AGNES

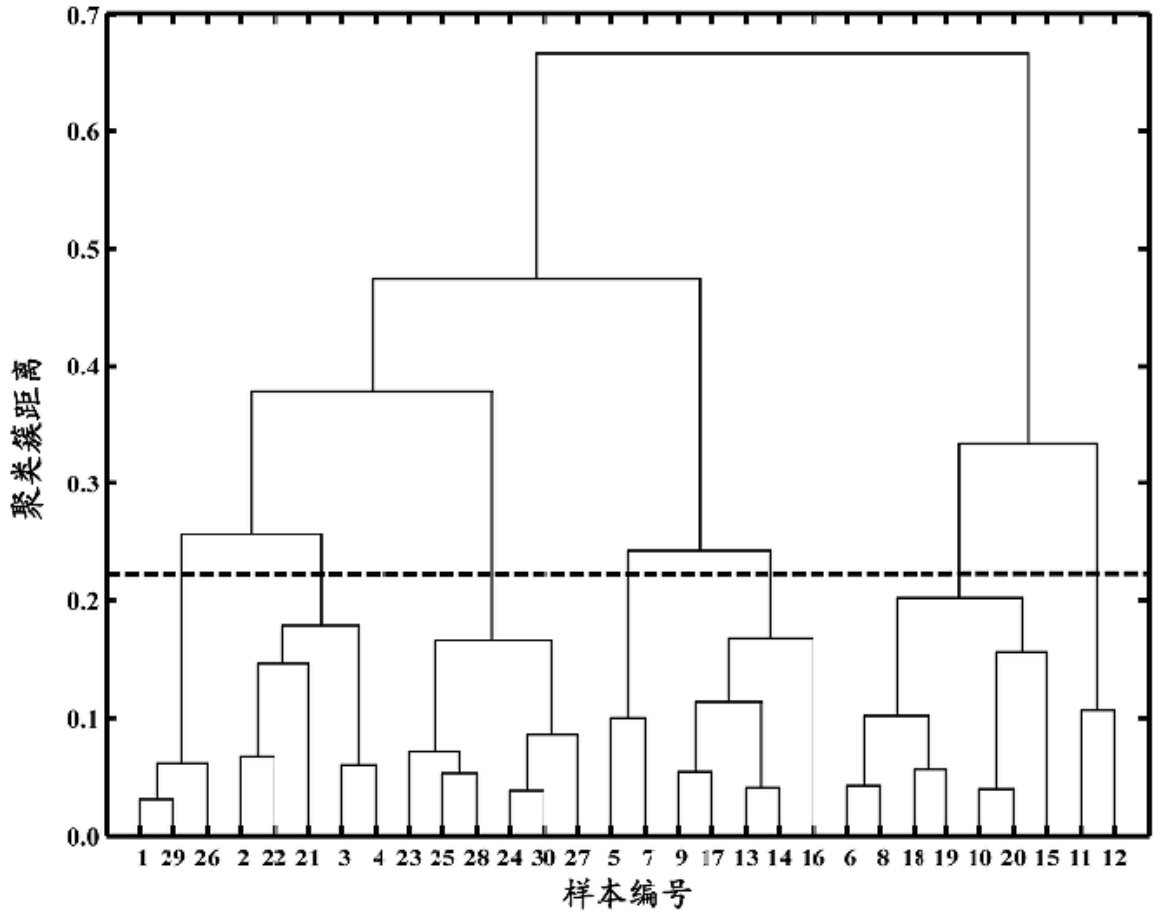


图 9.12 西瓜数据集 4.0 上 AGNES 算法生成的树状图(采用 $d_{\max}$). 横轴对应于样本编号, 纵轴对应于聚类簇距离.

大纲

- 相关概念
- 距离度量
- 聚类准则
- 聚类方法
- 聚类评价
- 前沿进展

聚类评价

南京大学 NANJING UNIVERSITY

❑ 可考虑用以下几个指标来评价聚类效果

- 聚类中心之间的距离
  - 距离值大，通常可考虑分为不同类
- 聚类域中的样本数目
  - 样本数目少且聚类中心距离远，可考虑是否为噪声
- 聚类域内样本的距离方差
  - 方差过大的样本可考虑是否属于这一类

❑ 讨论

- 聚类目前还没有一种通用的准则，往往需要根据实际应用来选择合适的方法

聚类评价

南京大学 NANJING UNIVERSITY

❑ 外部指标 (external index)

- 将聚类结果与某个“参考模型”(reference model) 进行比较
- 如 Jaccard 系数，FM 指数，Rand 指数

❑ 内部指标 (internal index)

- 直接考察聚类结果而不用任何参考模型
- 如 DB 指数，Dunn 指数等

❑ 基本想法

- “簇内相似度” （intra-cluster similarity）高，并且
- “簇间相似度” （inter-cluster similarity）低

聚类评价

南京大学 NANJING UNIVERSITY

❑ 常用评价指标（标签未知）

- **Compactness (CP) 紧密度**

$$\overline{CP}_i = \frac{1}{|\Omega_i|} \sum_{x_i \in \Omega_i} \|x_i - w_i\| \qquad \overline{CP} = \frac{1}{K} \sum_{k=1}^{K} \overline{CP}_k$$

- Ω_i表示聚类得到的一个簇，w_i表示该簇的中心，K表示簇（类）的个数。$\overline{\text{CP}}$值越小表示类内越紧凑
- 缺点：没有考虑类间聚类效果

聚类评价

南京大学 NANJING UNIVERSITY

❑ 常用评价指标（标签未知）

- **Separation (SP) 间隔度**

$$\overline{SP} = \frac{2}{k^2 - k} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \|w_i - w_j\|_2$$

- w_i表示第i簇（类）的中心，w_j表示第j簇（类）的中心。$\overline{SP}$值越大表示类间越分散
- 缺点：没有考虑类内聚类效果

聚类评价

南京大学 NANJING UNIVERSITY

❑ 常用评价指标（标签未知）

- **Davies-Bouldin Index (DBI) 戴维森堡丁指数/分类适确性指标**

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\overline{C_i}+\overline{C_j}}{\|w_i - w_j\|_2}\right)$$

- $\overline{C_i}$表示第i簇（类）的紧密度，w_i表示第i簇（类）的中心。$\boldsymbol{DB}$值越小，表示类内越紧凑，类间越分散
- 缺点：使用欧式距离，对于环状分布聚类评价很差

聚类评价

南京大学 NANJING UNIVERSITY

❑ 常用评价指标（标签未知）

- **Dunn Validity Index (DVI) 邓恩指数**

$$DVI = \frac{\min\limits_{0<m\neq n<K}\left\{\min\limits_{\substack{\forall x_i\in\Omega_m\\ \forall x_j\in\Omega_n}}\{\|x_i - x_j\|\}\right\}}{\max\limits_{0<m\leq K}\max\limits_{\forall x_i,x_j\in\Omega_m}\{\|x_i - x_j\|\}}$$

- x_i表示簇Ω_m中第i个样本， x_j表示簇Ω_n中第j个样本，计算任意两个簇元素的最短距离（类间）除以任意簇中的最大距离（类内）
- 缺点：对离散点的聚类测评很高、对环状分布测评效果差

聚类评价

南京大学 NANJING UNIVERSITY

❑ 常用评价指标（**标签已知**）

- Cluster Accuracy (CA) 聚类准确率
- Rand index (RI) 兰德指数
- Adjusted Rand index (ARI) 调整兰德指数
- Mutual Information (MI) 互信息
- Normalized Mutual Information (NMI) 归一化互信息

聚类评价

南京大学 NANJING UNIVERSITY

❑ 推荐阅读

Fahad A, Alshatri N, Tari Z, et al. **A survey of clustering algorithms for big data: Taxonomy and empirical analysis**[J]. IEEE transactions on emerging topics in computing, 2014, 2(3): 267-279.

大纲

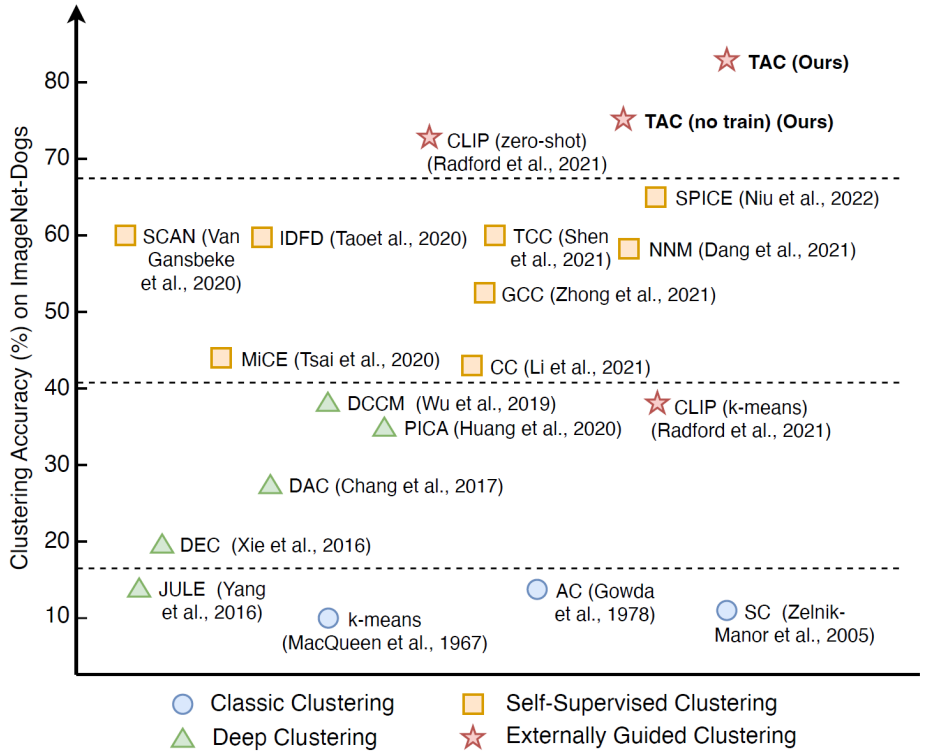
- 相关概念
- 距离度量
- 聚类准则
- 聚类方法
- 聚类评价
- 前沿进展

前沿进展：外部知识指导的聚类

南京大学 NANJING UNIVERSITY

❑ 聚类算法的演进：从内部监督到外部指导

- **传统聚类方法**：基于数据分布假设(k-means, 谱聚类)
- **深度聚类方法**：利用神经网络提取更有辨别力的特征
- **自监督聚类方法**：通过数据增强和动量策略构建监督信号
- **外部知识指导聚类(新范式)**：利用外部知识构建监督信号



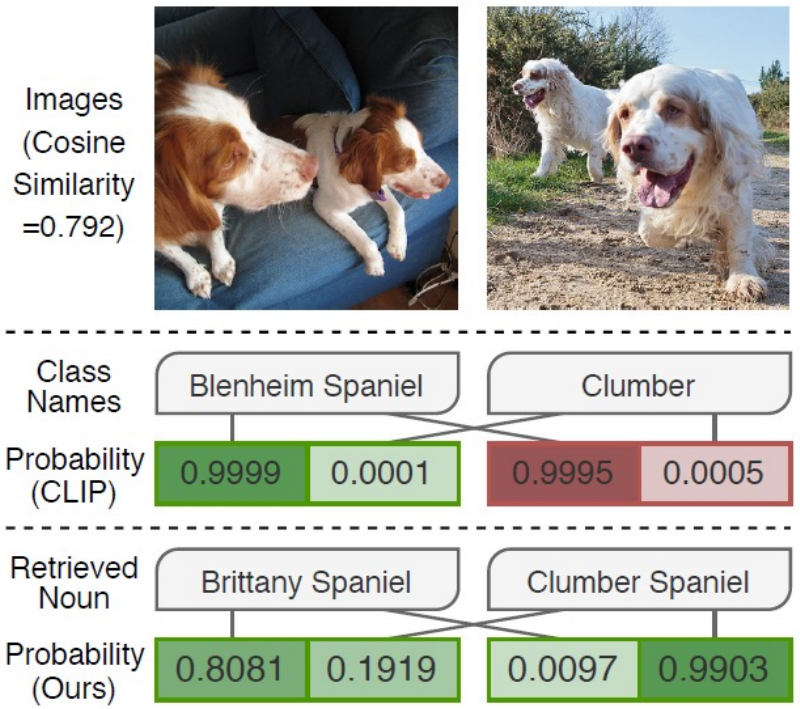
Yunfan Li, et al. Image Clustering with External Guidance. ICML 2024.

前沿进展：外部知识指导的聚类

❑ TAC: 文本辅助的图像聚类方法

- ✓ 如何构建文本空间（无类别名称先验的情况下）
- ✓ 如何协同文本和图像模态进行聚类

- 利用k-means聚类找到图像语义中心
- 将WordNet中的所有名词分类到这些语义中心
- 选择最具辨别力的名词构建文本空间
- 为每个图像检索相应的文本对应物

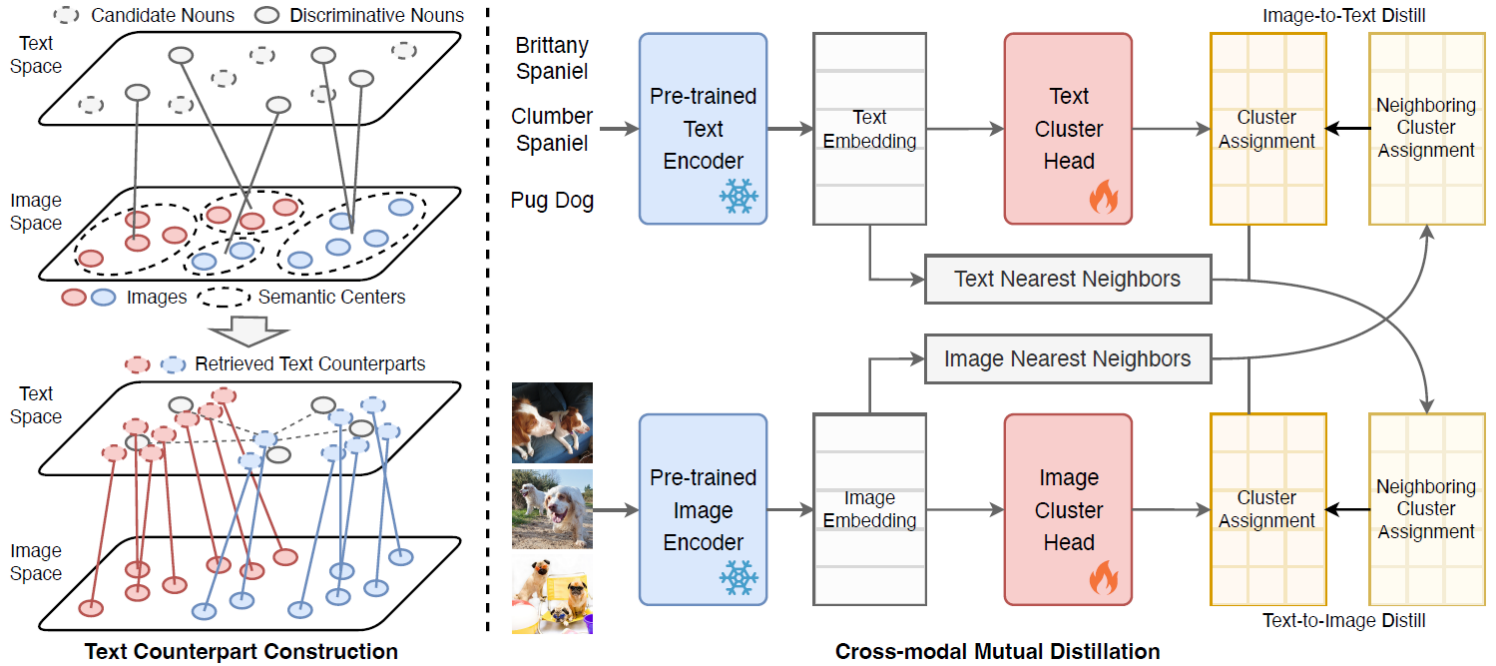


Yunfan Li, et al. Image Clustering with External Guidance. ICML 2024.

前沿进展：外部知识指导的聚类

❑ TAC的跨模态互蒸馏策略

- 图像→文本蒸馏: 鼓励文本具有与其对应图像的近邻一致的聚类分配
- 文本→图像蒸馏: 鼓励图像具有与其文本对应物的近邻一致的聚类分配



Yunfan Li, et al. Image Clustering with External Guidance. ICML 2024.

前沿进展：公平聚类的鲁棒性挑战

南京大学 NANJING UNIVERSITY

❑ 聚类问题中的公平性挑战

- **监督学习 vs. 聚类**：在有监督学习中，可以通过标签评估不同群体的错误率差异；而在聚类中，没有这样直接的评估机制
- **隐式偏见**：聚类算法可能会无意中将相似的数据点分到不同组，或将不同的数据点强制归为一组，导致对某些保护群体的不公平结果
- **评估困难**：缺乏标签使得评估聚类结果的公平性更加复杂

❑ 现有公平聚类方法的关注点

- 如何定义聚类结果的公平性
- 如何调整算法以提高聚类结果的公平性
- 如何在保持聚类质量的同时实现公平性

公平聚类算法对对抗性攻击的脆弱性

⬇

研究空白

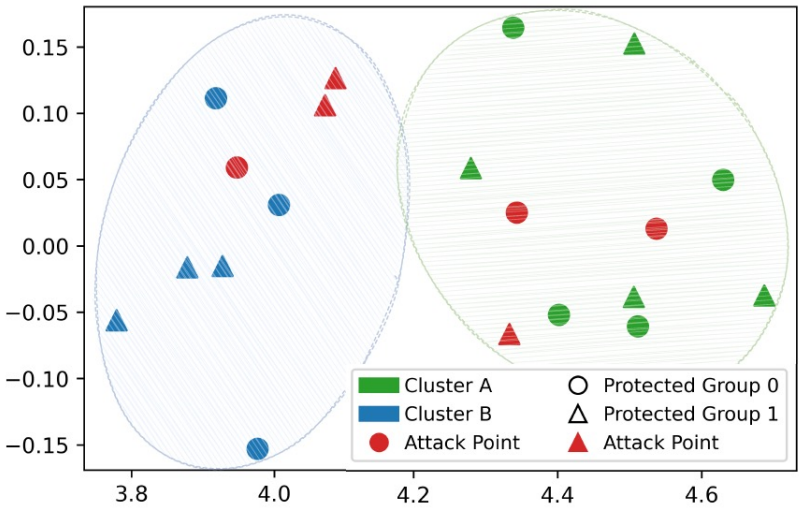
Anshuman Chhabra, et al. Robust Fair Clustering: A Novel Fairness Attack and Defense Framework. ICLR 2023.

前沿进展：公平聚类的鲁棒性挑战

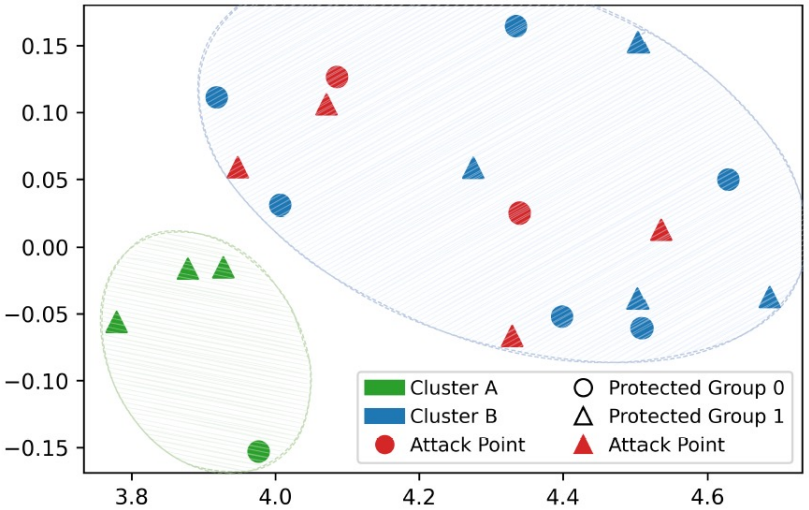
南京大学 NANJING UNIVERSITY

❑ 研究问题

- 公平聚类算法是否容易受到降低公平性的对抗性攻击?
- 如何开发对此类攻击具有鲁棒性的公平聚类模型?



(a) Pre-attack



(b) Post-attack

Anshuman Chhabra, et al. Robust Fair Clustering: A Novel Fairness Attack and Defense Framework. ICLR 2023.

前沿进展：公平聚类的鲁棒性挑战

南京大学 NANJING UNIVERSITY

❑ 防御策略：共识公平聚类 (CFC)

- **两阶段防御机制**：

  (1) 通过随机子集生成多个基础聚类，构建共现矩阵

  (2) 基于图嵌入和公平约束学习鲁棒聚类表示

- **优化目标**：

  包含自监督对比损失、公平聚类损失和结构保持损失

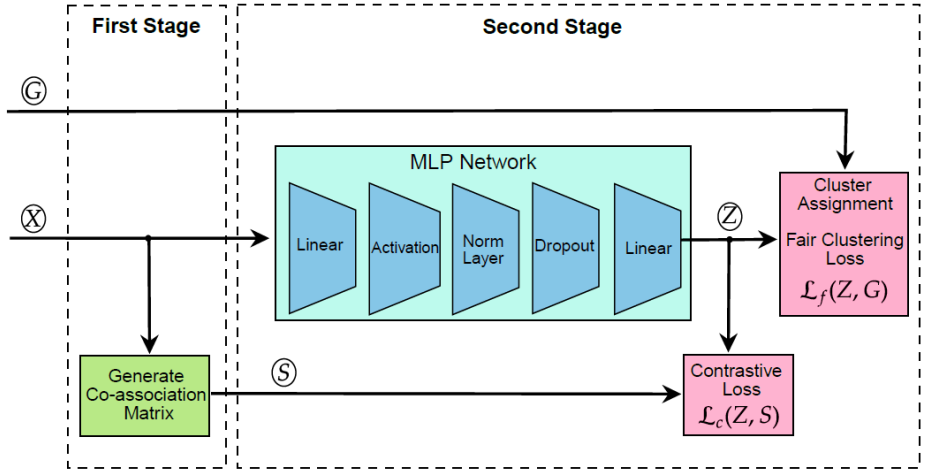


Figure 3: Our proposed CFC framework.

Anshuman Chhabra, et al. Robust Fair Clustering: A Novel Fairness Attack and Defense Framework. ICLR 2023.

无监督学习实践内容

作业要求：参考“无监督学习实践内容-自主练习”完成思考题目

提交要求：不作为小作业，自行练习为主

负责助教：刘尚格

答疑邮箱： lshangge@smail.nju.edu.cn

提交邮箱： ~~nju_ml@163.com~~

提交时间： ~~2025年*月**日晚24:00~~

南京大学 NANJING UNIVERSITY

谢 谢！

联系方式：liwenbin@nju.edu.cn

更多信息：www.liwenbin.cn