

无监督学习

高 阳，李文斌

<http://cs.nju.edu.cn/rl>, 2021.03.23



物以類聚
人以群分

大纲

聚类相关概念

距离度量

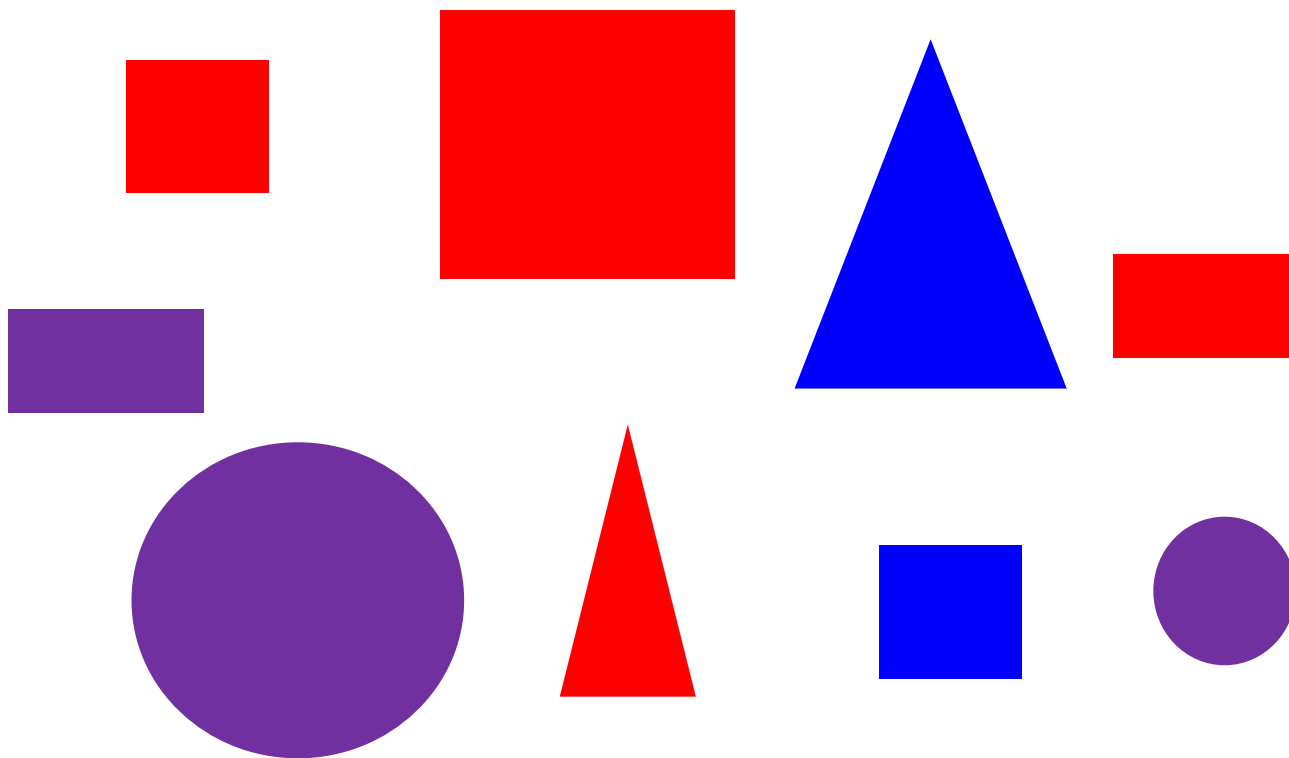
聚类准则

聚类方法

聚类评价

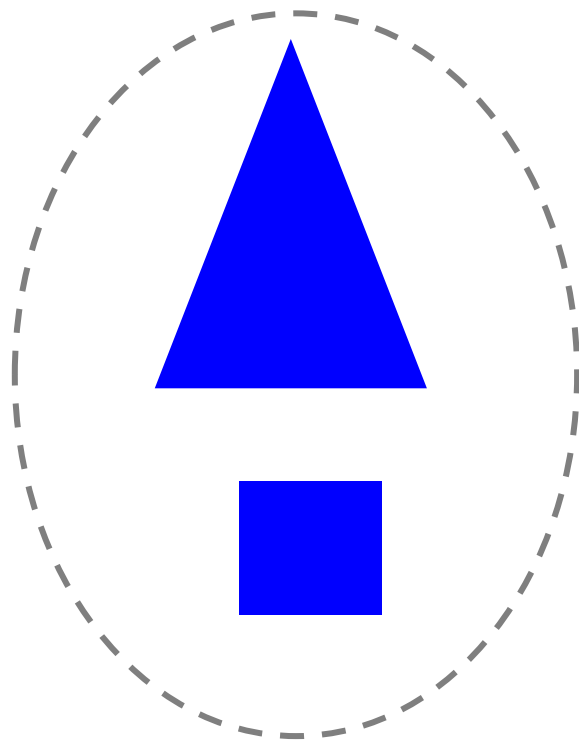
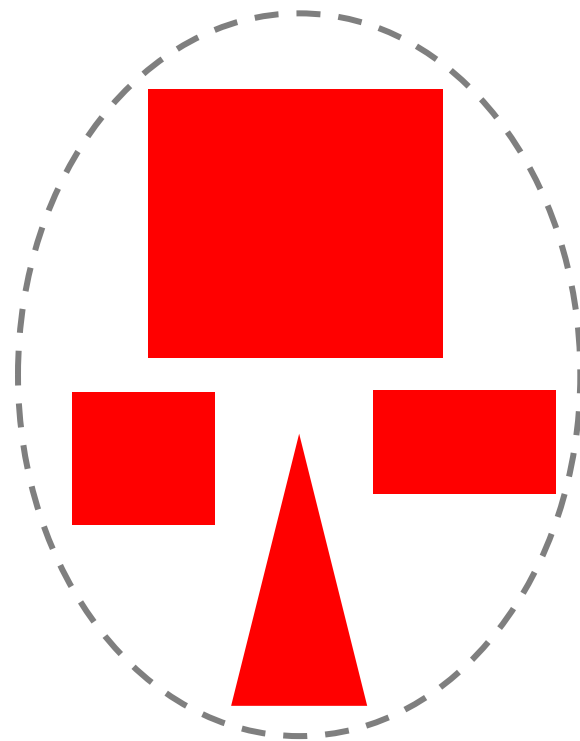
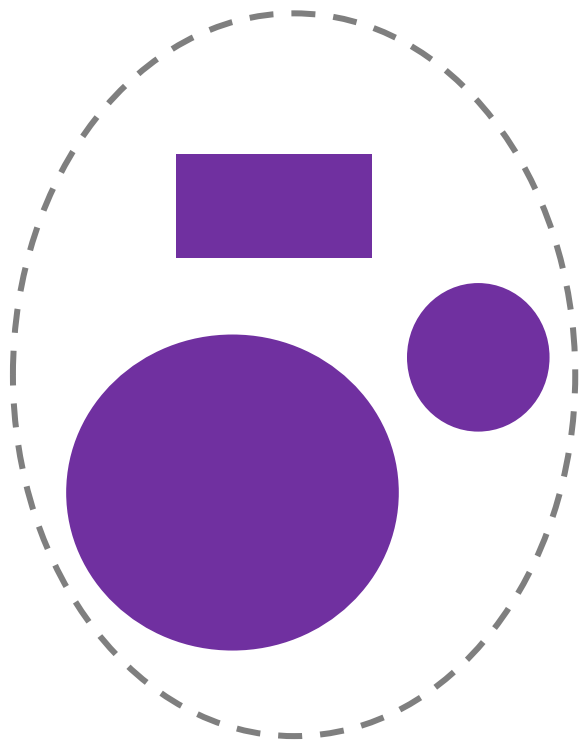
一个作业

- 为下面图形进行分类（**幼儿园作业**）



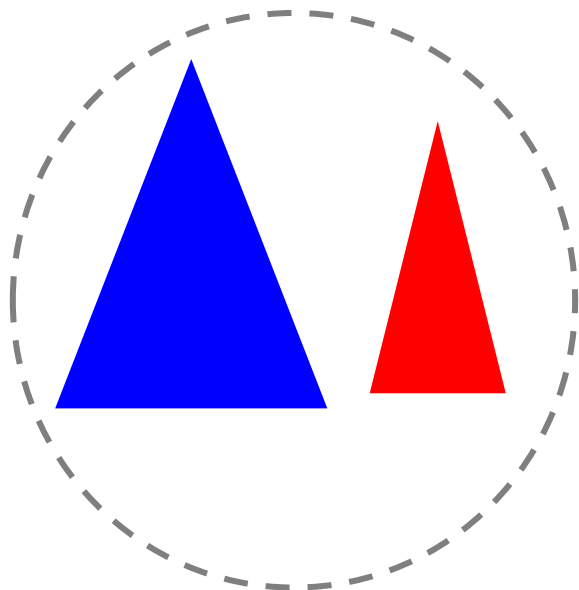
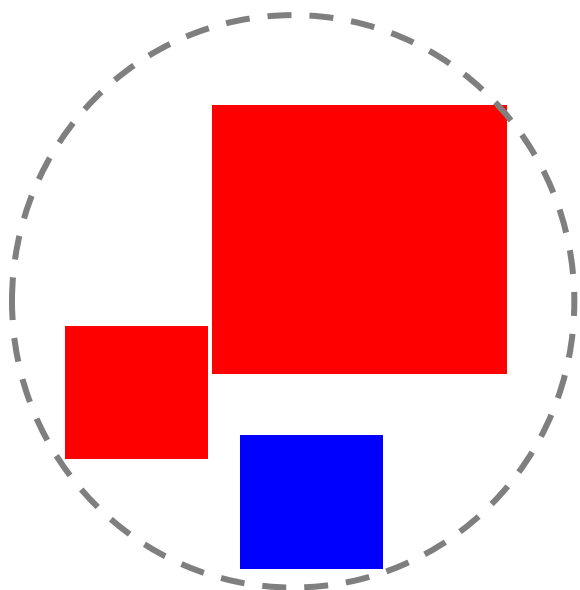
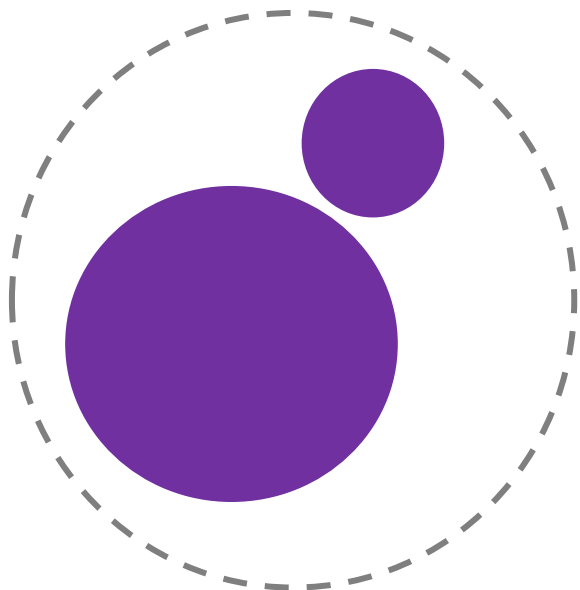
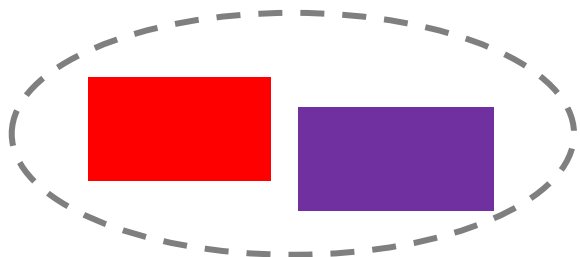
一个作业

- 颜色



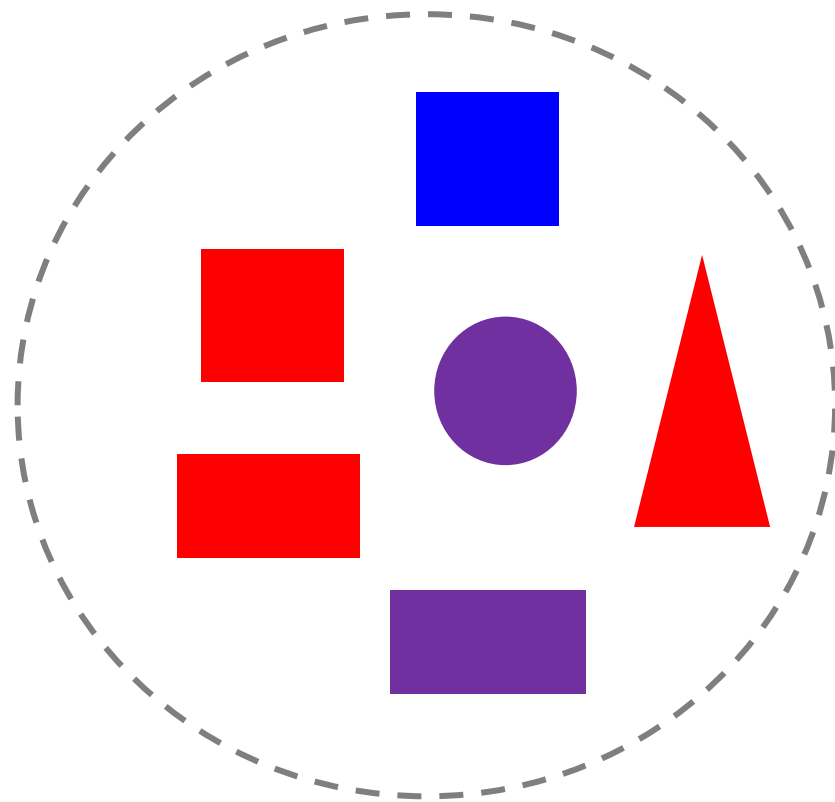
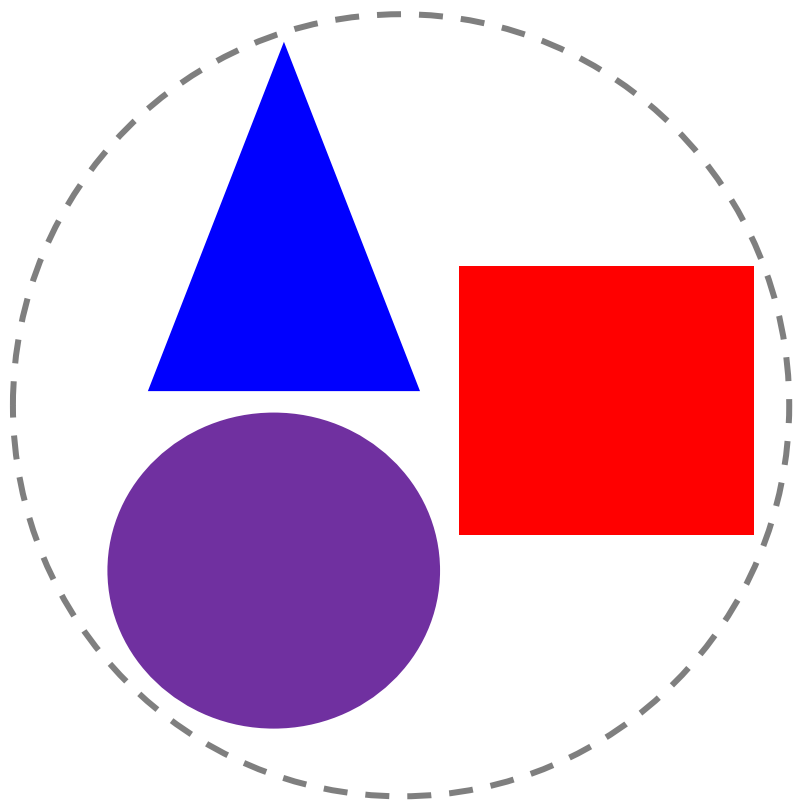
一个作业

- 形状



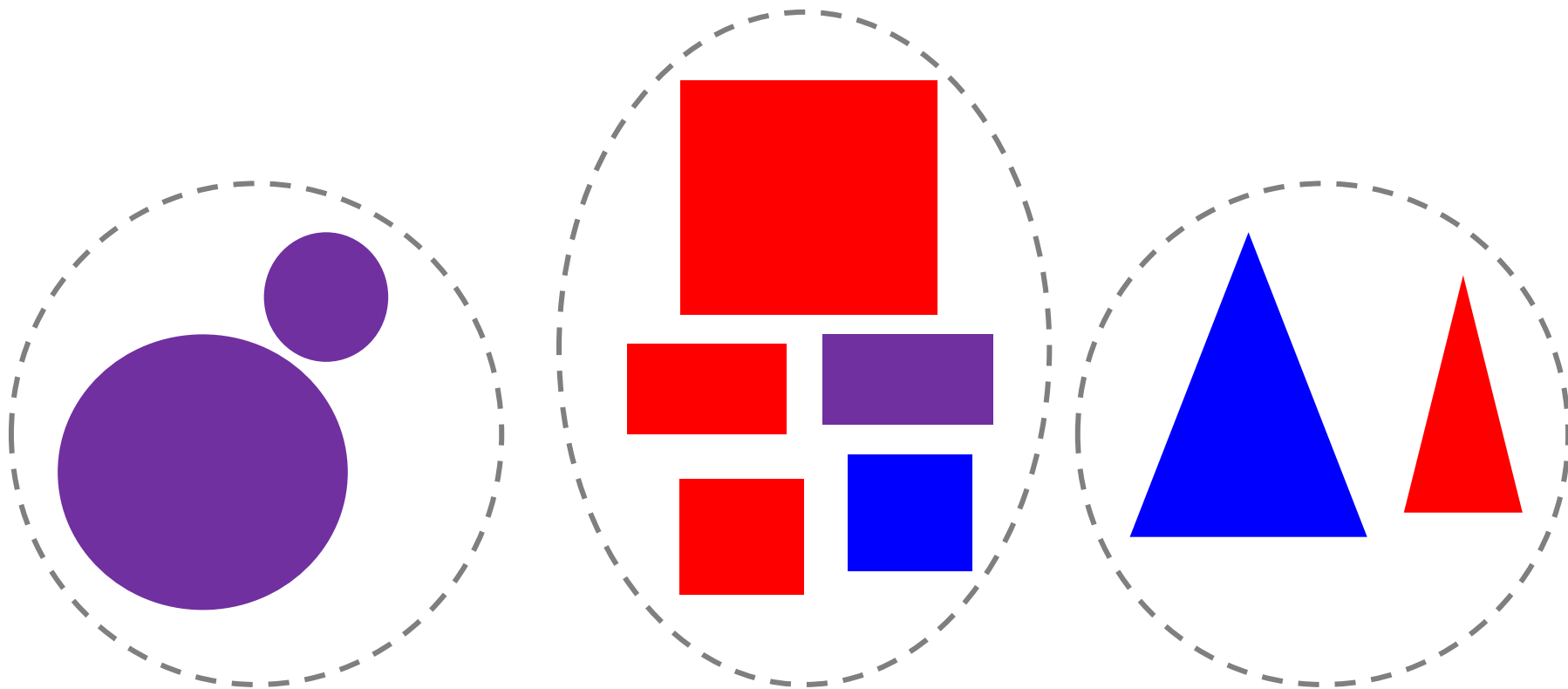
一个作业

- 大小



一个作业

- 顶点数



必须记住

○ ○ ○

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“标准”，使以往算法对它无能为力

大纲

聚类相关概念

距离度量

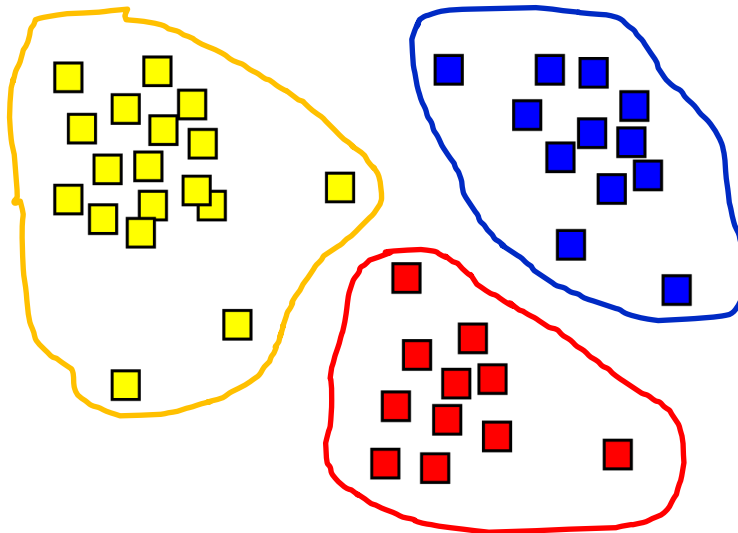
聚类准则

聚类方法

聚类评价

相关概念

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



相关概念

- 聚类算法：

根据给定的相似性评价标准，将一个数据集合并组/划分成几个聚类（簇）

- 数学形式化：

样本集合： $D = \{x_1, x_2, \dots, x_m\}, x_i \in \mathbb{R}^d$

聚类成 k 个簇： $\{C_l | l = 1, 2, \dots, k\}$

$$C_{l'} \cap_{l' \neq l} C_l = \emptyset$$

$$D = \bigcup_{l=1}^k C_l$$

相关概念

- 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些点，**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）进行模式分类（聚类）的。

相关概念

- 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

- 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

- 特征对聚类的影响

羊	狗
蓝鲨	蜥蜴
毒蛇	猫
麻雀	海鸥
金鱼	蛙

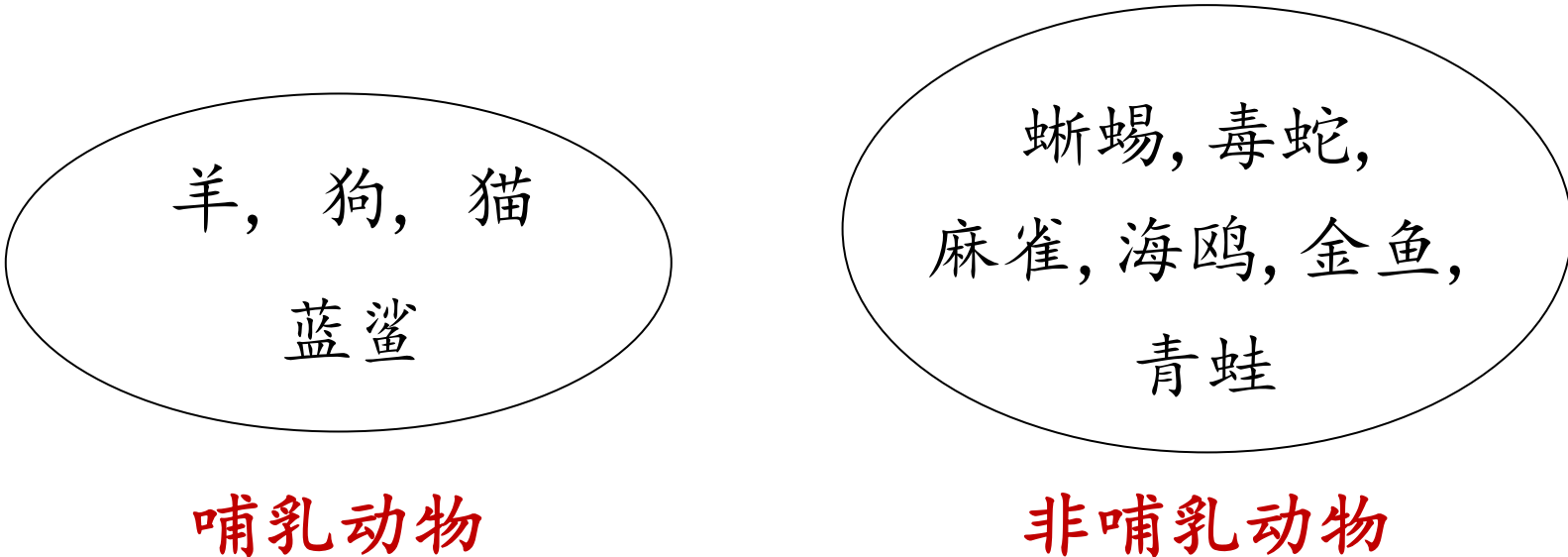
要对这些动物进行聚类，不同的特征则有不同的分法

相关概念

- 特征对聚类的影响

羊 狗 蓝鲨 蜥蜴 毒蛇
猫 麻雀 海鸥 金鱼 蛙

按繁衍后代的方式分：



羊, 狗, 猫
蓝鲨

哺乳动物

蜥蜴, 毒蛇,
麻雀, 海鸥, 金鱼,
青蛙

非哺乳动物

相关概念

- 特征对聚类的影响

羊 狗 蓝鲨 蜥蜴 毒蛇

猫 麻雀 海鸥 金鱼 蛙

按是否存在肺分：

按生活环境分呢？

金鱼，蓝鲨

无肺

羊，狗，猫，
蜥蜴，毒蛇，麻雀，
海鸥，青蛙

有肺

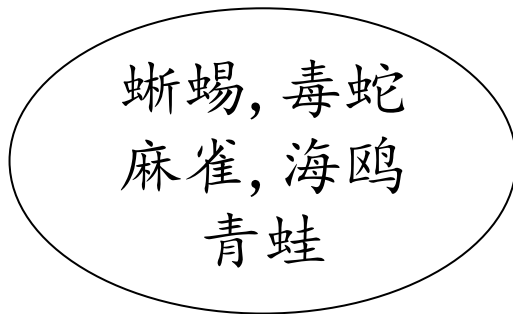
相关概念

- 特征对聚类的影响

羊 狗 蓝鲨 蜥蜴 毒蛇

猫 麻雀 海鸥 金鱼 蛙

按繁衍后代方式和肺是否存在分：



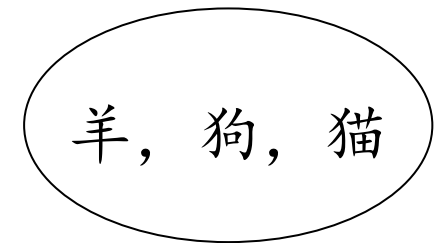
非哺乳且有肺



非哺乳且无肺



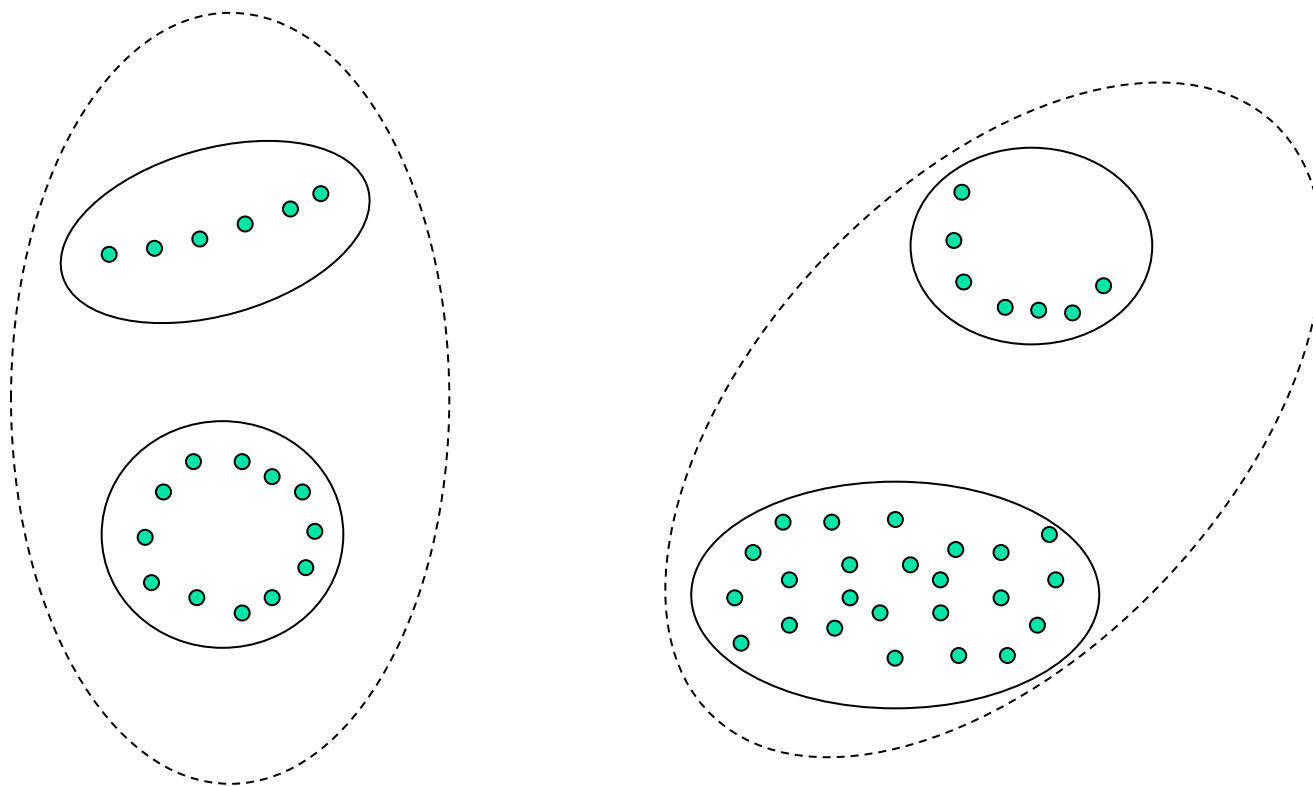
哺乳且无肺



哺乳且有肺

相关概念

- 距离度量对聚类的影响



数据的粗聚类是2类,细聚类为4类

相关概念

- 聚类的关键
 - 特征的选取或设计
 - 距离度量函数的选择

相关概念

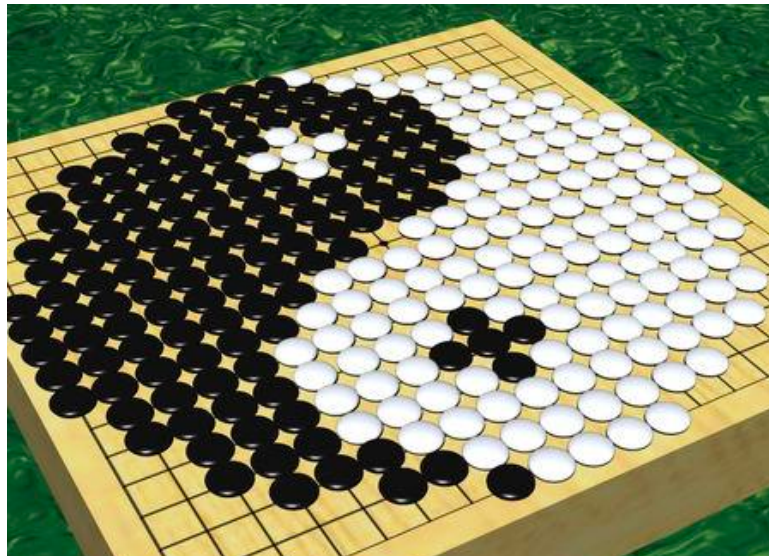
- 聚类分析的有效性

聚类分析方法是否有效，与**数据分布**形式有很大关系！

- 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。

相关概念

- 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

聚类相关概念

距离度量

聚类准则

聚类方法

聚类评价

距离度量

- 目的

度量同类样本间的相似性和不同类样本间的差异性

距离度量

- 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 x, y 映射到一个非负实数，并满足以下四条公理：

1. 非负性 $d(x, y) \geq 0$

2. 唯一性 $d(x, y) = 0 \leftrightarrow x = y$

3. 对称性 $d(x, y) = d(y, x)$

4. 三角不等式 $x, y, z \in X, d(x, z) \leq d(x, y) + d(y, z)$

距离度量

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

欧氏距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)}$$

余弦相似性

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$$

曼哈顿距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_1$$

切比雪夫距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_\infty$$

马氏距离

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)}$$

距离度量

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

欧氏距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

曼哈顿距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_1$$

切比雪夫距离

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_\infty$$



闵可夫斯基距离
Minkowski distance

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

大纲

聚类相关概念

距离度量

聚类准则

聚类方法

聚类评价

聚类准则

- 类的定义

- 类的定义有很多种，类的划分具有人为规定性，这反映在定义的选取及参数的选择上。一个聚类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 x_i 与 x_j 间的距离有

$$d(x_i, x_j) \leq h$$

- 其中 h 为给定的阈值，称 S 对于阈值 h 组成一类。

聚类准则

- 聚类有了模式的相似性度量/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

聚类准则

- 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

- 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 $\{\mathbf{x}\}$ 和类别 $\{S_j, j = 1, 2, \dots, c\}$ 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

- 聚类准则函数方法

- 一种聚类准则函数J的定义

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

- J代表了属于 c 个聚类类别的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

聚类相关概念

距离度量

聚类准则

聚类方法

聚类评价

聚类方法

- 基于试探的聚类搜索算法
- 系统聚类法
- 动态聚类法

基于试探的聚类搜索算法

- 按最近邻规则的简单试探法

给定N个待分类的数据样本 $\{x_1, x_2, \dots, x_N\}$ ，要求按距离阈值T，将它们分类到聚类中心 z_1, z_2, \dots

- 给定N个待分类的数据样本 $\{x_1, x_2, \dots, x_N\}$ ，要求按距离阈值T，将它们分类到聚类中心 z_1, z_2, \dots

第一步：任取一样本 x_i 作为一个聚类中心的初始值，例如令 $z_1 = x_1$

计算 $D_{21} = \|x_2 - z_1\|_2$

若 $D_{21} > T$ ，则确定一个新的聚类中心 $z_2 = x_2$

否则 x_2 属于以 z_1 为中心的聚类

第二步：假设已有聚类中心 z_1, z_2

计算 $D_{31} = \|x_3 - z_1\|_2$

$D_{32} = \|x_3 - z_2\|_2$

若 $D_{31} > T$ 且 $D_{32} > T$ ，则得一个新的聚类中心 $z_3 = x_3$

否则 x_3 属于离 z_1 和 z_2 中的最近者

.....

如此重复下去，直至将 N 个模式样本分类完毕。

基于试探的聚类搜索算法

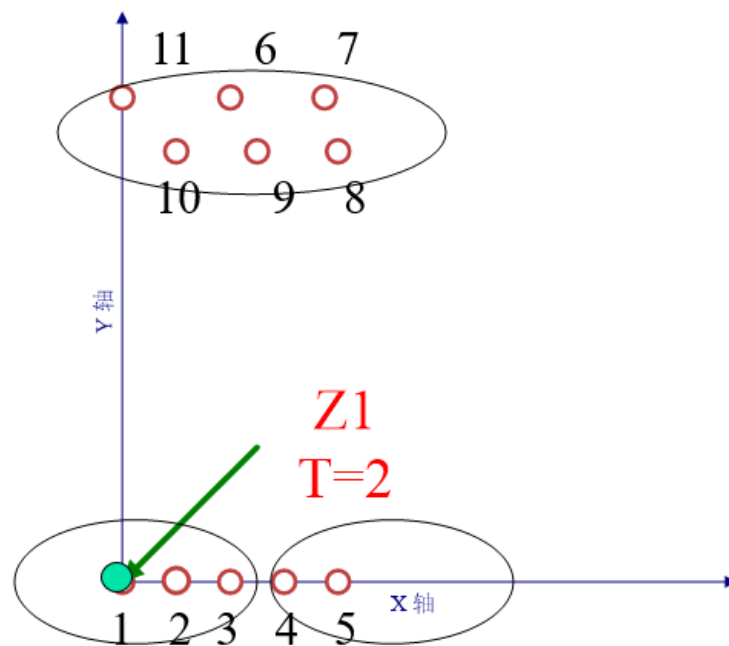
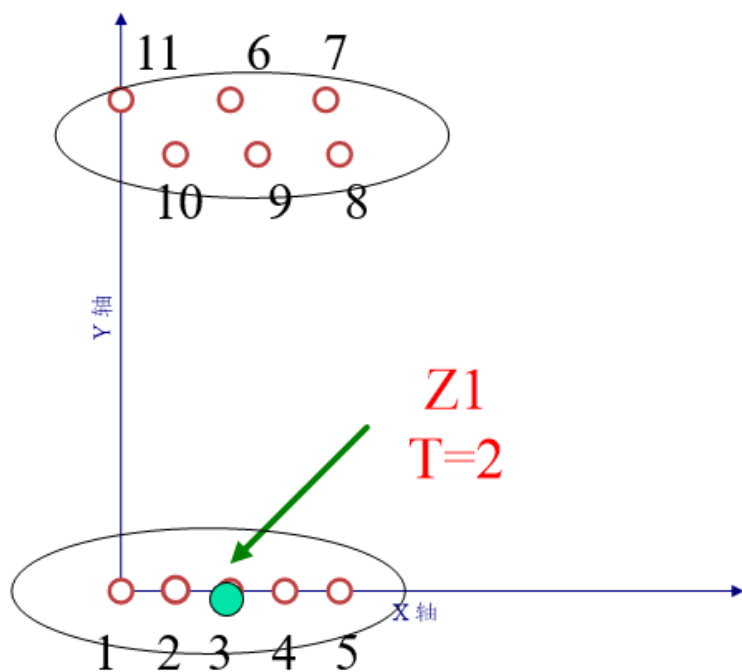
- 按最近邻规则的简单试探法

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
- 待分类样本的排列次序
- 距离阈值 T 的大小
- 样本分布的几何性质

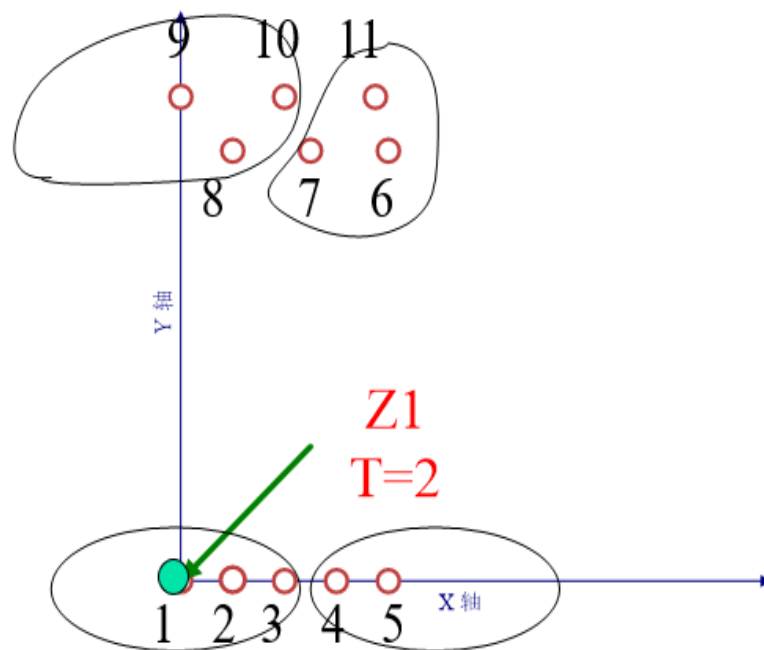
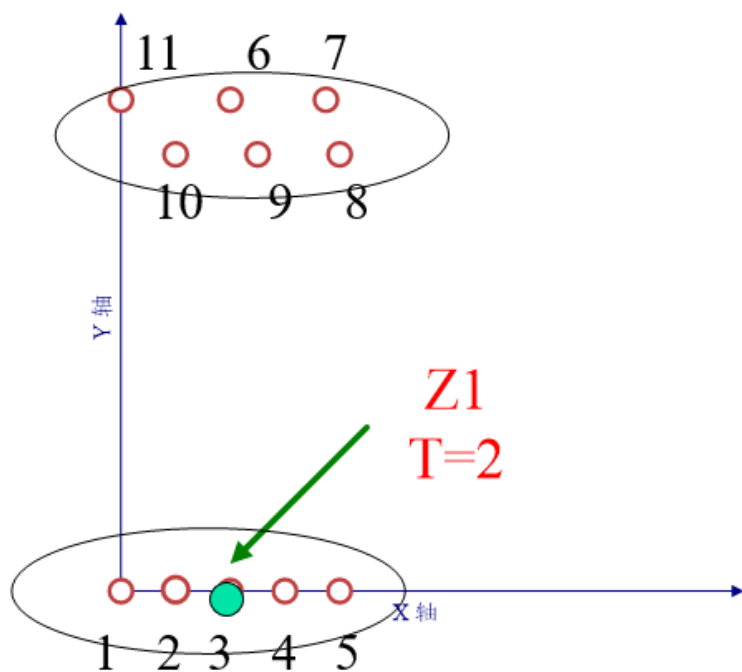
基于试探的聚类搜索算法

初始点不同



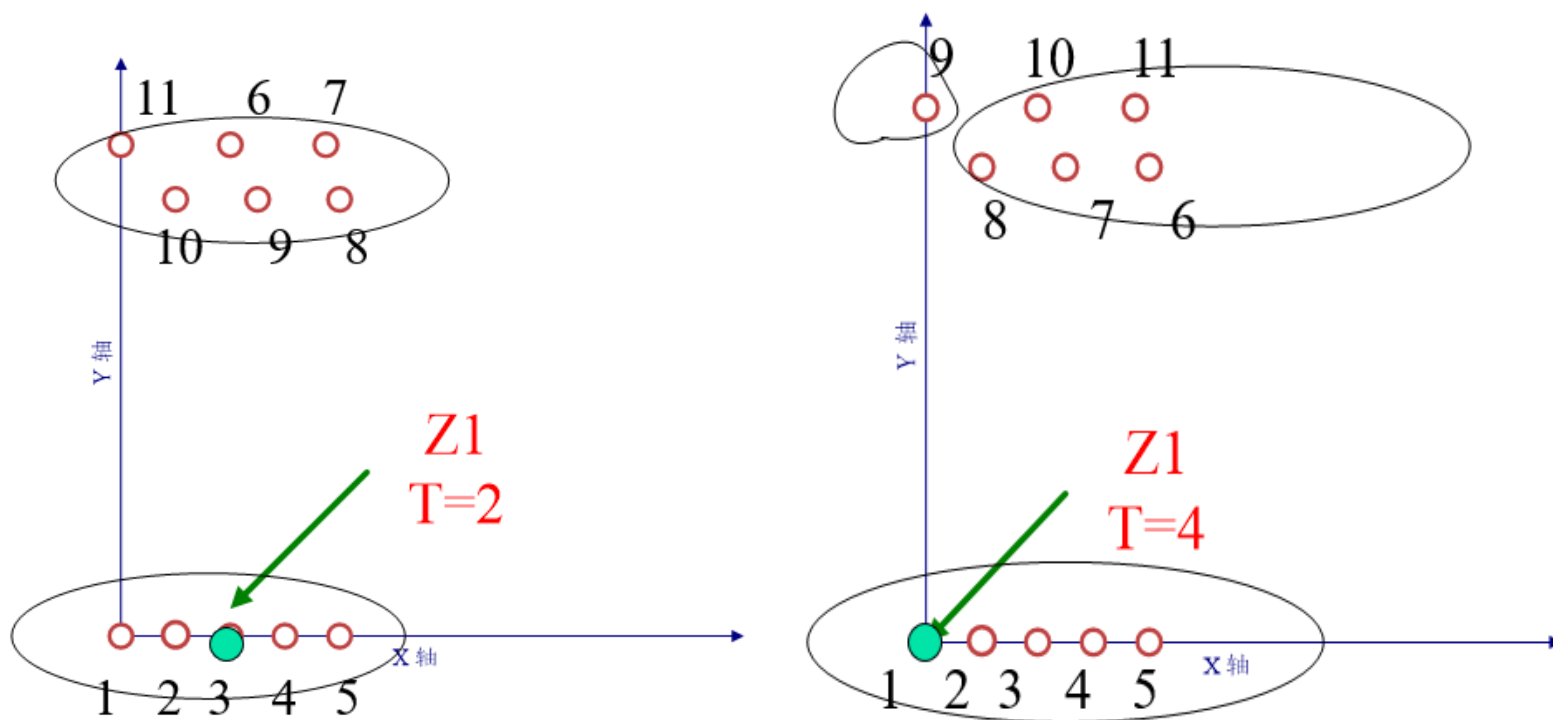
基于试探的聚类搜索算法

样本次序不同



基于试探的聚类搜索算法

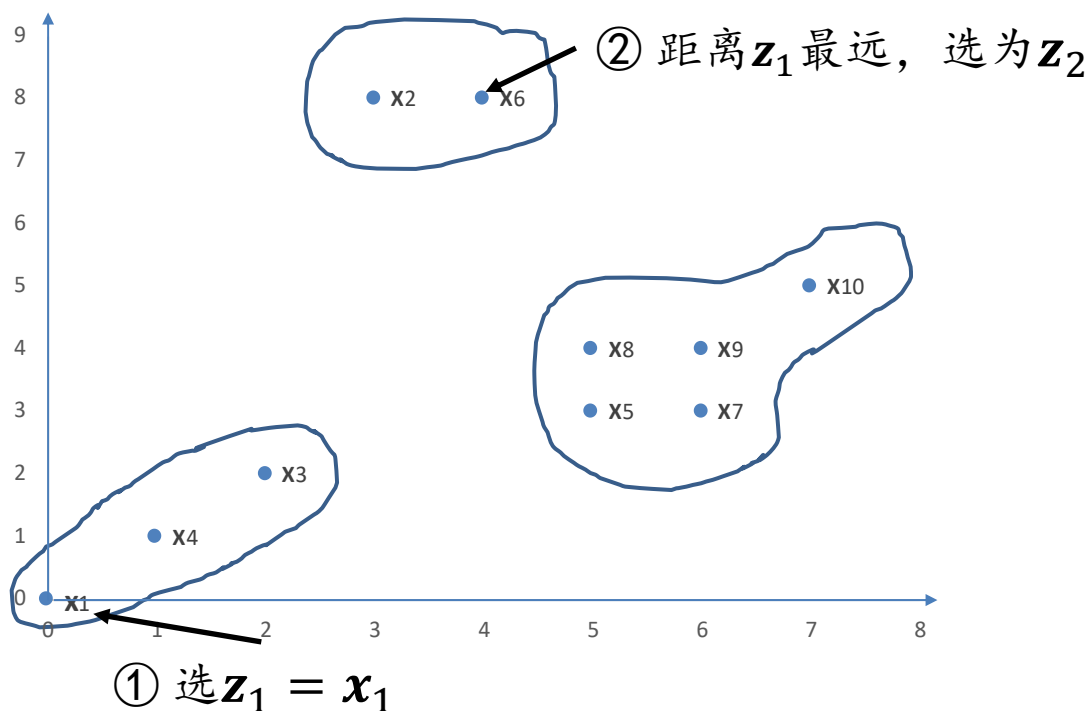
阈值T不同



基于试探的聚类搜索算法

- 最大最小距离算法

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。



第一步：选任意一个样本作为第一个聚类中心，如 $z_1 = x_1$

第二步：选距离 z_1 最远的样本作为第二个聚类中心。

经计算， $\|x_6 - z_1\|$ 最大，所以 $z_2 = x_6$

第三步：逐个计算各样本 $\{x_i, i = 1, 2, \dots, N\}$ 与 $\{z_1, z_2\}$ 之间的距离，即

$$D_{i1} = \|x_i - z_1\|$$

$$D_{i2} = \|x_i - z_2\|$$

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

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 $\|z_1 - z_2\|$ 的一定比例以上，则相应的样本点取为第三个聚类中心 z_3 ，即

若 $\max\{\min(D_{i1}, D_{i2}), i = 1, 2, \dots, N\} > \theta \|z_1 - z_2\|$ ，则 $z_3 = x_i$

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 $1/2$ 。

在此例中，当 $i=7$ 时，符合上述条件，故 $z_3 = x_7$

第五步：若有 z_3 存在，则计算 $\max\{\min(D_{i1}, D_{i2}, D_{i3}), i = 1, 2, \dots, N\}$ 。若该值超过 $\|z_1 - z_2\|$ 的一定比例，则存在 z_4 ，否则找聚类中心的过程结束。

在此例中，无 z_4 满足条件。

第六步：将模式样本 $\{x_i, i = 1, 2, \dots, N\}$ 按最近距离分到最近的聚类中心：

$z_1 = x_1$: $\{x_1, x_3, x_4\}$ 为第一类

$z_2 = x_6$: $\{x_2, x_6\}$ 为第二类

$z_3 = x_7$: $\{x_5, x_7, x_8, x_9, x_{10}\}$ 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。

系统聚类法

- 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

• 算法流程

第一步：设初始模式样本共有 N 个，每个样本自成一类，即建立 N 类， $G_1^{(0)}, G_2^{(0)}, \dots, G_N^{(0)}$ 。计算各类之间的距离（初始时即为各样本间的距离），得到一个 $N \times N$ 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： $G_1^{(n+1)}, G_2^{(n+1)}, \dots$ 。

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 $G_1^{(n+1)}, G_2^{(n+1)}, \dots$ 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。）

系统聚类法

- 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）

动态聚类法

- 基本思想

- 首先选择若干个样本点作为聚类中心，再按某种聚类准则（通常采用最小距离准则）使样本点向各中心聚集，从而得到初始聚类；
- 然后判断初始分类是否合理，若不合理，则修改聚类
- 如此反复进行修改聚类的迭代算法，直至合理为止。

- 代表算法

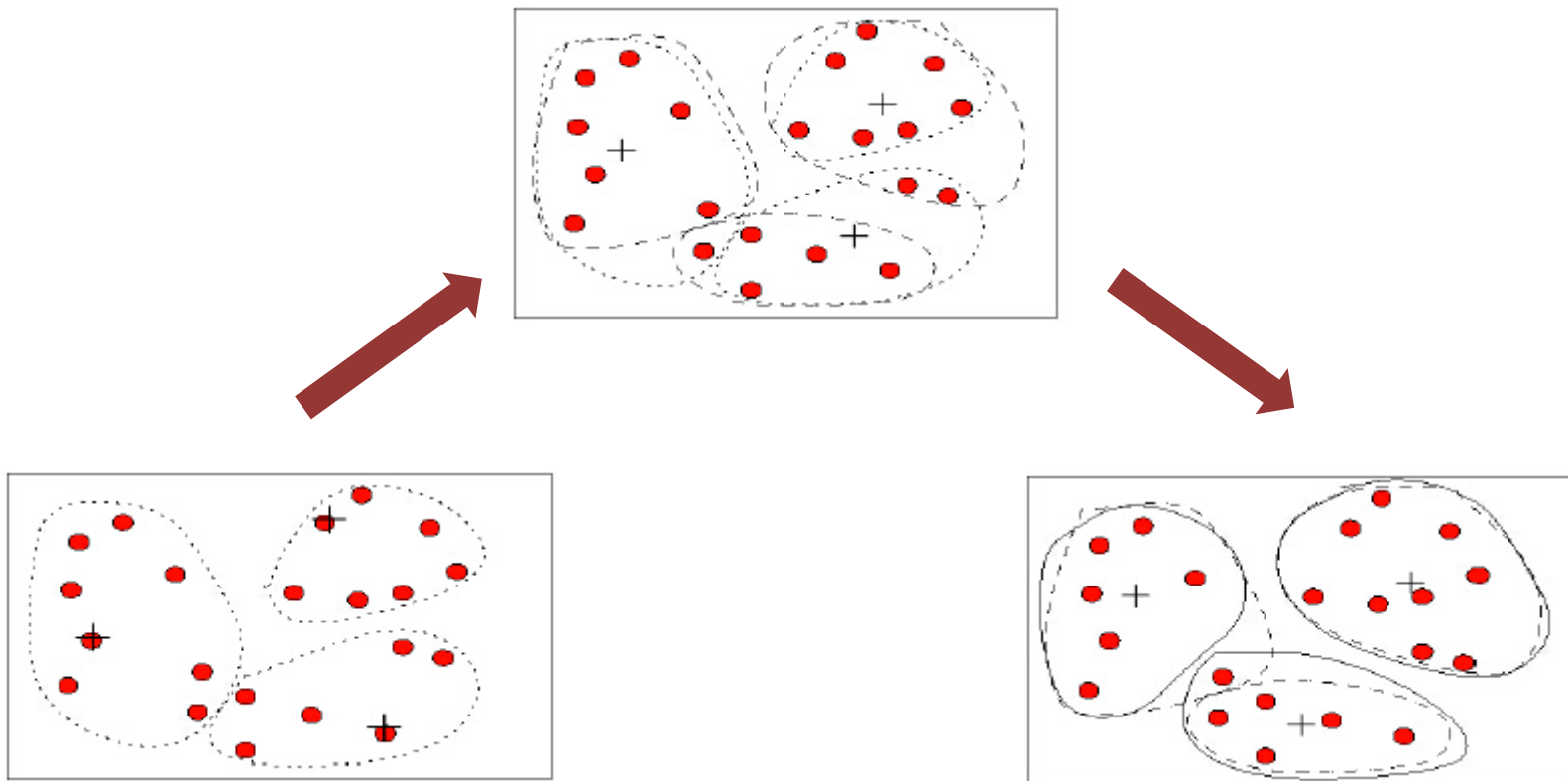
- K-means算法
- ISODATA算法（迭代自组织数据分析算法）

K-means算法

• 算法流程

- Step1: 选择一个聚类数量 k
- Step2: 初始化聚类中心 μ_1, \dots, μ_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 (使用某种距离度量方法), 将样本点分距离它最近的聚类中心所属的聚类
- Step4: 重新计算聚类中心, 聚类中心为属于这一个聚类的所有样本的均值
- Step5: 如果没有发生样本所属的聚类改变的情况, 则退出, 否则, 返回Step3继续。

K-means算法



K-means算法

- 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
 - 模式样本的几何性质
 - 。 。 。
- 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

- 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 c_1 ;
Step 2: 首先计算每个样本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 $D(x)$ 表示; 接着计算每个样本被选为下一个聚类中心的概率 $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$ 。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

- 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

运行过程中能够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 $C = \{c_1, c_2, \dots, c_{K_0}\}$ ；
Step 2: 针对数据集中每个样本 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 c_i ，重新计算它的聚类中心 $c_i = \frac{1}{ c_i } \sum_{x \in c_i} x$ (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

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：。

$$m_{new} = \frac{1}{n_i + n_j} (n_i m_i + n_j m_j).$$

上式中的 n_i 和 n_j 表示这两个类别中的样本个数，新的聚类中心可以看作是对这两个类别进行加权求和。如果其中一个类所包含的样本个数较多，所合成的新类就会更加偏向它。。

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 $K = K + 1$ 。 $m_i^{(+)} = m_i + \sigma_{max}, m_i^{(-)} = m_i - \sigma_{max}。$

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

- **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。

大纲

聚类相关概念

距离度量

聚类准则

聚类方法

聚类评价

聚类评价

- 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

- 常用评价指标（**标签未知**）

- ✓ Compactness (CP) 紧密度

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

Ω_i 表示聚类的到的一个簇， w_i 表示该簇的中心， K 表示簇（类）的个数。
 \overline{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

- 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

- 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

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\bar{C}_i + \bar{C}_j}{\|w_i - w_j\|_2} \right)$$

\bar{C}_i 表示第*i*簇（类）的紧密度， w_i 表示第*i*簇（类）的中心

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

- 常用评价指标（**标签未知**）

- ✓ **Dunn Validity Index (DVI) 邓恩指数**

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

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）

DVI值越大，表示类内越紧凑，类间越分散

- 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

- 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

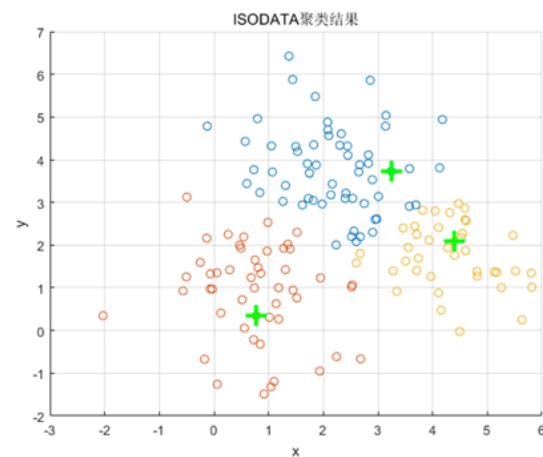
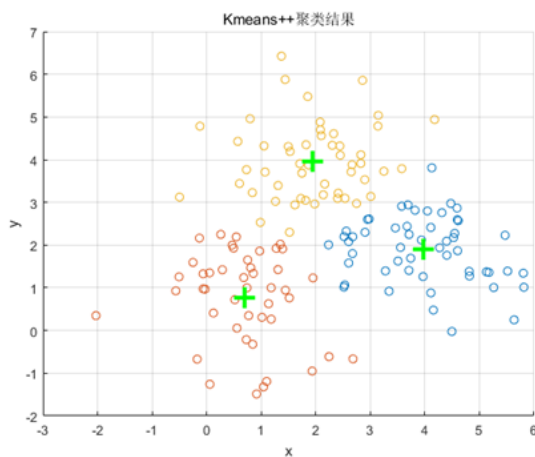
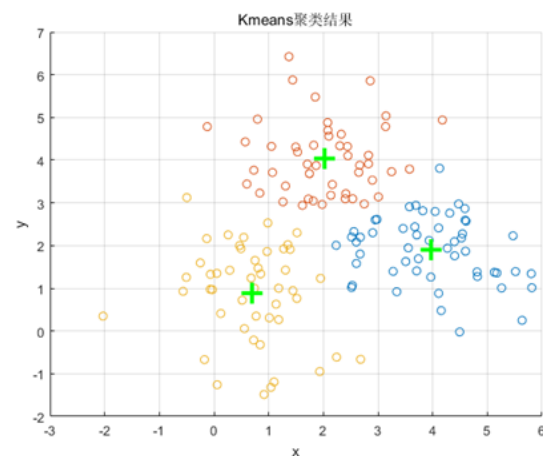
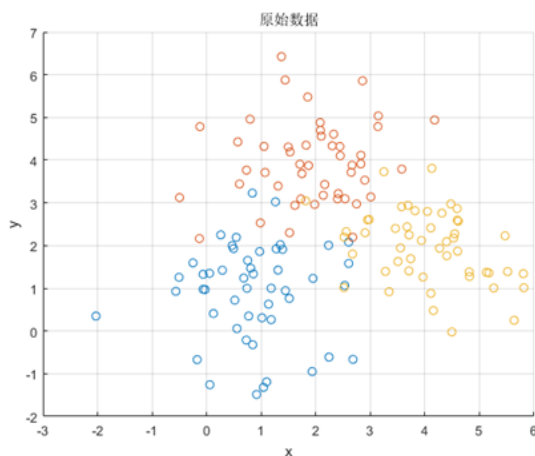
聚类评价

- 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.

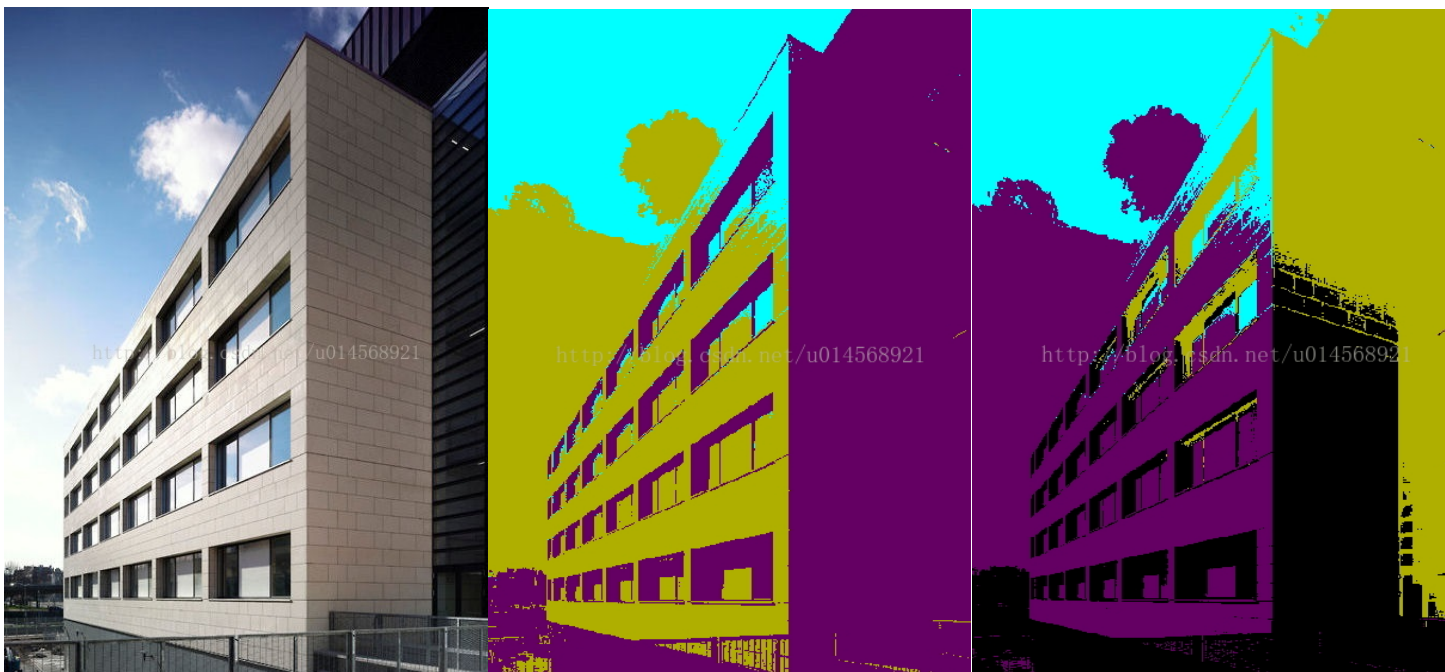
聚类评价

- 可视化：二维高斯分布的数据聚类结果



聚类评价

- 可视化（图像分割）



谢谢！