

无监督学习和聚类

翟婷婷

扬州大学
信息工程（人工智能）学院
zhdt@yzu.edu.cn

2023春

课程目标

- 了解无监督学习的学习动机，与有监督学习的差别。
- 掌握数据聚类的基本概念、特点、应用领域、聚类的一般流程。
- 理解常用的聚类算法的原理，包括
 - 基于划分的聚类：k均值聚类和k中心点聚类
 - 基于层次的聚类：合并式聚类和分裂式聚类
 - 基于密度的聚类：DBSCAN并了解每种聚类方法的优势和局限性。
- 能够编程实现k均值聚类、k中心点聚类、以及DBSCAN聚类，用于解决具体的模式识别问题。

有监督学习vs.无监督学习

➤ 有监督学习 (supervised learning):

- ✿ 依赖于已经标注好类别标签的样本构成的训练集
- ✿ 旨在从训练集中学习到具体的决策规则
- ✿ 常用于：分类、回归

➤ 无监督学习 (unsupervised learning):

- ✿ 训练集中样本的类别标记未知
- ✿ 旨在发现训练集中内在的结构或规律
- ✿ 常用于：聚类、概率密度估计

无监督学习的动机

① 减小对数据进行类别标注的代价:

有监督学习依赖于大量**标注的训练数据**。对数据进行标注，往往是通过**人工**方式，将大量的数据打上标签，分好类，从而形成供分类器学习的训练集。

数据标注工作是一个**劳力密集型**工作，需要耗费大量的人力资源。无监督学习方法，例如聚类，可以自动发现数据中的分组，**为人工标注数据提供辅助**，从而减少人工标注数据的代价。

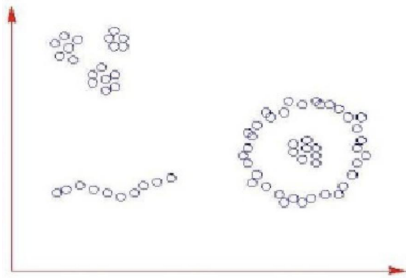
② 在任何探索性的工作中，无监督方法可以揭示观测数据的一些内部结构和规律，为分类器设计提供依据。

聚类分析(clustering analysis)

➤ 聚类是一种最常见的无监督学习。

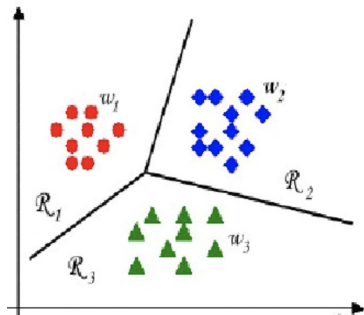
聚类就是把样本集中的样本按照**相似的程度**划分成不同的集合，每个集合称为一个“簇”(cluster)，使得**同一个簇内样本相似程度**远大于**不同簇的样本的相似程度**。

物以类聚，人以群分：相似物分在一起，而同一群人可能有着相似的性格爱好，共同目标等。

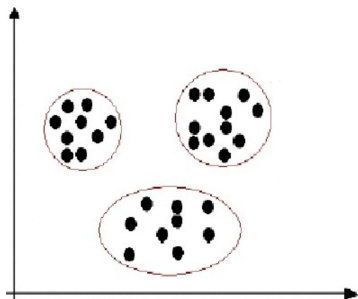


聚类分析(clustering analysis)

➤ 分类与聚类的区别:



给定已标记的训练集，构造分类决策边界，对特征空间进行划分。



给定一组数据，发现数据中潜在的结构或分组。

聚类分析的用途

- 聚类是从数据中学习类别划分，因此聚类的基本功能就是去主动挖掘数据中隐藏的知识和结构，解释样本间的内在联系；
- 聚类以数据间的内在关联为依据，自动对庞杂的数据进行整理，形成良好的数据组织结构，为后续的数据利用奠定良好的基础；
- 聚类能够实现样本集样本的初始划分，对样本进行自动类别标注，为学习分类器准备好初始数据，这也是“无监督学习”为“有监督学习”做出的贡献；
- 聚类也常用于样本集的简化，即通过聚类，将相似度比较高的样本进行合并或删减，或用典型的样本来代替非典型的样本，以大幅度减少样本集中的样本数量，降低问题求解的复杂度。

聚类分析(clustering analysis)

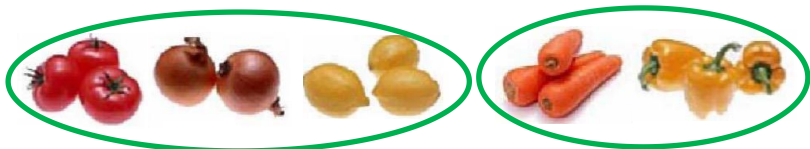
- 聚类的基本假设：同一个簇内部样本的相似程度远大于不同簇的样本的相似程度。
- 聚类的依据是：样本的相似程度，可以用相似性度量函数来衡量，例如：欧式距离、余弦相似度等。
- 聚类可以分为：硬聚类和软聚类
 - 硬聚类：只允许每个样本被划分到一个簇中
 - 软聚类：允许每个样本在不同程度上隶属于不同的簇
- 聚类的特点：
 - ✿ 聚类结果受聚类准则的影响
 - ✿ 聚类结果受相似性度量函数的选择的影响
 - ✿ 聚类结果受各个特征的量纲标尺的影响

聚类分析(clustering analysis)

- 对蔬菜进行聚类：聚类结果受**聚类准则**的影响



- 按照蔬菜**形状**进行聚类：

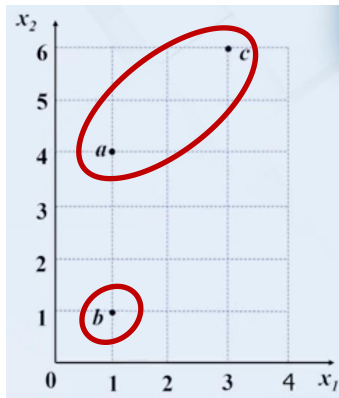


- 按照蔬菜**颜色**进行聚类：

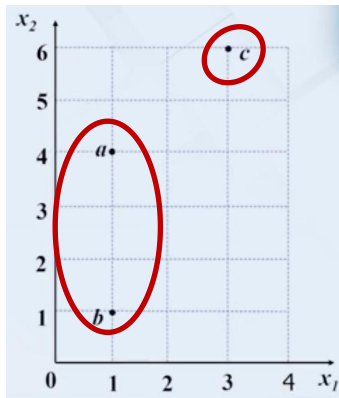


聚类分析(clustering analysis)

➤ 聚类结果受相似性度量函数的影响



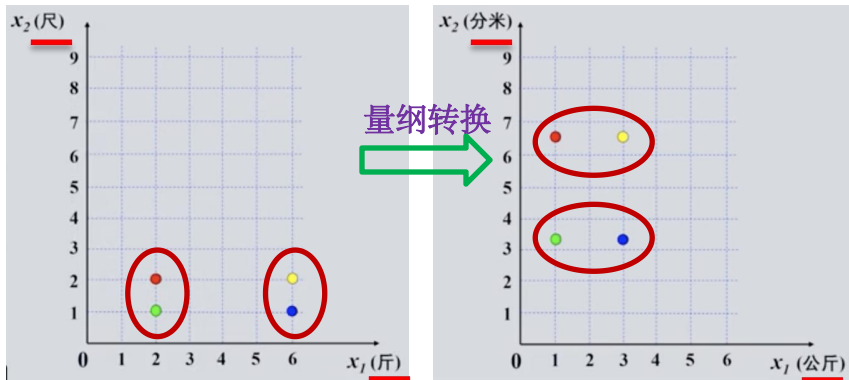
三个样本采用欧氏距离聚成两类



三个样本采用曼哈顿距离聚成两类

聚类分析(clustering analysis)

- 聚类结果受各个特征的量纲标尺的影响



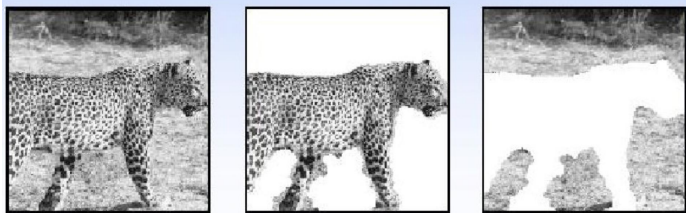
四个样本采用欧氏距离聚成两类

聚类分析(clustering analysis)

- 量纲标尺对聚类结果的影响，是由于在不同的量纲标尺下**特征取值的大小**出现了差异，在计算相似度时，不同维度的特征被赋予了不同的权重，即**取值越大的特征维度，在相似度计算中影响就越大**。
- 除非是由于模式识别任务自身的要求，我们人为赋予不同特征不同的权重，否则需要在进行数据预处理的过程中，消除这种量纲标尺带来的不良影响。
- **特征归一化**：min-max标准化、z-score标准化、排名归一化等。
- 是否进行量纲尺度的标准化，要根据样本集的具体情况决定。例如在某些任务中，某些特征确实应该具有比其他特征更大的权重，此时不宜进行归一化。

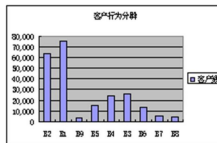
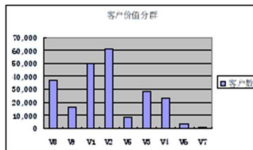
聚类的应用领域

- 在图像处理领域可用于对前景图像和背景图像的分割:



聚类的应用领域

- 在**经济领域**可用于对客户进行分类，发现最优价值的客户群;



- 在**信息检索领域**可用于合并相似检索结果，减少检索返回量;

<http://www.fyqiang.net/SinWen/SinWenTag.asp?ID=2159&TAG=浙江网>

高考: 武汉市五千余名考生今日赴考 武汉市积极做好2005年高考考务工 浙江全面贯彻落实
安化环讯会议精神 武汉市土地整理 市司法局平遥镇新向四川地震灾区捐款 市委组织
新桥镇组织募捐支援四川地震 武汉市发起为爱心捐助灾区活动 武汉市开展“...
www.fyqiang.net/SinWen/SinWenTag.asp?ID=215 ... 125K 2008-5-25 - 百度快照

北华人 北海新闻 news.bhnharen.cn

用户名: 密码: 自动登录 隐身登录 注册 ... 2008-06-12 昨日上午, 广西首批赴四川地震灾区
损失较大防疫卫生监督工作任务的 ... 广西30万考生今日迅速高考考场 全区考点达241个 2
008-06-7 6月7日, 8日 2008年广西高考如期举行。记者从自治区 ...
news.buzhouren.cn/news_d/bzhsharen/index ... 125K 2008-6-15 - 百度快照

房天下 FunSou.com-中国网上看房产第一网 房产在出租 房产中

做了12年的房产经纪到平, 后续部分居民已拿到证 (2008年6月20日 09:43) 康山6.49万
考生周末中考 考生数量比... 大马人民捐款地震灾区已达3500万林吉特 (2008年6月20
日 6:16) 交通事故发布台湾游客及家属今日返台 (2008年6月...
www.0591tc.com/news/china.asp 125K 2008-6-20 - 百度快照

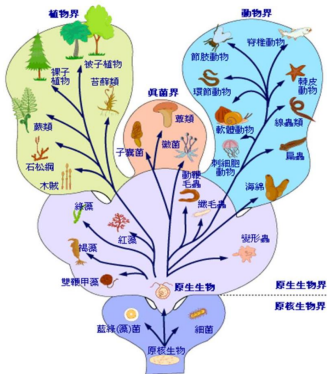
提示: 为了提供相关的结果, 我们省略了一些内容相似的条目, [点击这里](#)可以看到所有搜索结果。

[上一頁](#) [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

相关搜索: 地震灾区高	地震灾区高寒时间	四川地震灾区高寒	甘肃地震灾区高寒
地震灾区	恰地震灾区的一封信	四川地震灾区	地震重灾区

聚类的应用领域

- 在生物领域可以用于基因分析和生物的分类;



- 在数据处理领域可以用于对数据进行自动标注, 从大量数据中挖掘知识, 或者进行数据集简化。

聚类的流程

- 完整的数据聚类过程一般包括以下步骤：



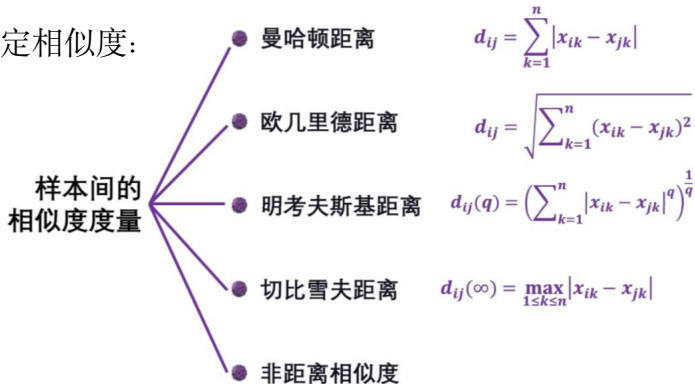
- 选定特征：选择哪些特征作为聚类特征来使用，会直接影响到聚类的结果。
- ✿ 首要考虑因素是聚类任务本身的需求，哪些特征是任务本身所关注的；
 - ✿ 其次，选择对聚类最有效的那些特征，要使得采用这些特征完成聚类后，聚类的结果比较理想；
 - ✿ 最后，还要考虑特征的数量和计算复杂度。

聚类的流程

➤ 完整的数据聚类过程一般包括以下步骤:



➤ 确定相似度:

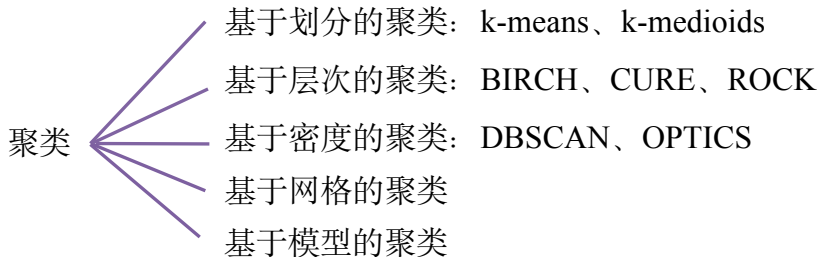


聚类的流程

- 完整的数据聚类过程一般包括以下步骤：



- 选择聚类算法：



聚类的流程

- 完整的数据聚类过程一般包括以下步骤：



- 评估聚类结果：

- ❁ 判断聚类结果是否达到了目标，无法通过已知的训练集来检验，需要定义一些评价指标来评估聚类效果。
- ❁ 评价指标要综合考虑簇内和簇间样本的相似度。好的聚类结果一定是簇内部相似度高，簇之间相似度低。
- ❁ 要优化评价指标，可以通过调整聚类算法的参数和过程，以达到更好的聚类效果。

典型的聚类算法

- 基于划分的聚类算法:

 - k -均值 (k-means)

 - k -中心点 (k-medoids)

- 基于层次的聚类算法:

 - 合并式聚类

 - 分裂式聚类

- 基于密度的聚类算法:

 - DBSCAN

基于划分的聚类算法

- 问题描述：给定一个样本集 $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^d$, 基于划分的聚类方法旨在将 n 个样本划分为 k 个集合 $S = \{S_1, S_2, \dots, S_k\}$, 使得

$$S_i \cap S_j = \emptyset, \quad \forall i \neq j$$

$$S_1 \cup S_2 \cup \dots \cup S_k = \mathcal{D}$$

- 典型的基于划分的算法:

k -均值 (k-means)

k -中心点 (k-medoids)

k -均值聚类算法

➤ 算法流程:

- ✧ 步骤1: 随机选择 k 个初始簇中心(聚类中心);
- ✧ 步骤2: 对所有样本, 分别计算其到 k 个簇中心的距离, 并将其划分入距离它最近的簇中心所在的簇中, 从而形成 k 个簇;
- ✧ 步骤3: 重新计算每个簇的均值向量作为新的簇中心;
- ✧ 步骤4: 如果簇中心没有任何改变, 算法停止, 否则回到步骤2。

k-均值聚类算法

➤ k-均值聚类算法:

距离: 平方欧式距离

簇中心: 该簇样本的均值向量

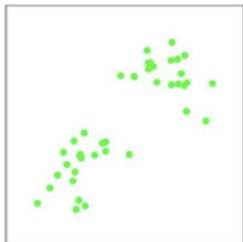
聚类规则: 将样本指派到距它最近的簇中心所在的簇中

收敛性: 可以证明K-Means一定会收敛, 且其聚类目标是**最小化误差平方和准则函数**, 即每个样本与其**所属的簇的均值向量**之间的误差的平方和:

$$J_e = \sum_{i=1}^k \sum_{x \in S_i} \|x - \mathbf{m}_i\|^2$$

其中, \mathbf{m}_i 是第 i 个簇 S_i 中样本的均值向量。 J_e 的值越小, 表明簇内样本的相似度越高。

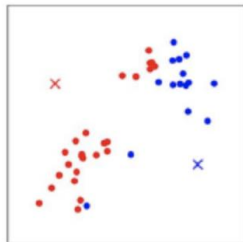
k -均值聚类过程



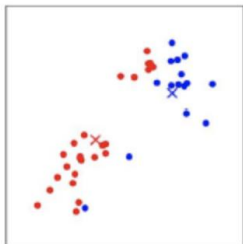
(a)



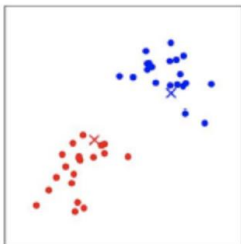
(b)



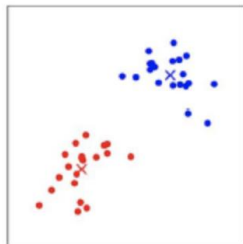
(c)



(d)



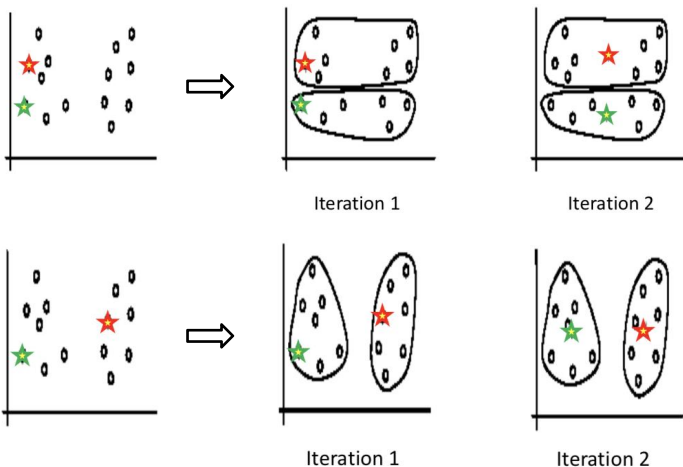
(e)



(f)

k-均值聚类：对初始簇中心敏感

- k -均值算法最终的聚类结果受初始簇中心的影响，因此聚类结果不唯一。

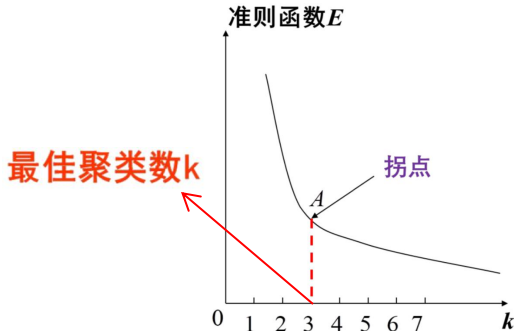


k-均值聚类：初始簇中心的选择

- 为了避免聚类结果受初始簇中心的影响，需要对初始簇中心进行特别选择。
- 常用的选择方法有以下几种：
 - ✳ 选择几何意义明显的特殊样本作为初始簇中心；
 - ✳ 选择距离最远的 k 个样本作为初始簇中心；
 - ✳ 先进行随机聚类，再将每个簇的均值向量作为初始簇中心；
 - ✳ 随机选择 k 个样本作为初始簇中心。

k-均值聚类：簇数 k 的确定

- k -均值算法需要预先指定要聚成的簇数 k ，但在许多情况下，并不知道样本集能够聚成几类。
- 一种方法：计算在不同的聚类簇数 k 下，聚类结果的准则函数值，取准则函数曲线的拐点 (Inflexion) 作为最佳的簇数 k 的值。



k-均值聚类的优点

➤ 原理简单，易于实现，收敛速度快。

➤ 算法高效：时间复杂度为 $O(tkn)$

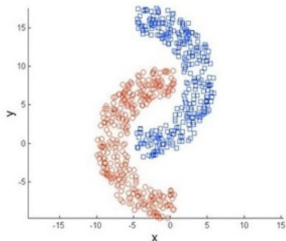
其中， n 是样本数， k 是簇的数目， t 是算法迭代次数。
 k 和 t 通常较小，所以 k -均值算法被认为一个线性时间的算法，即时间复杂度近似为 $O(n)$ 。

➤ 算法的可解释性比较好。

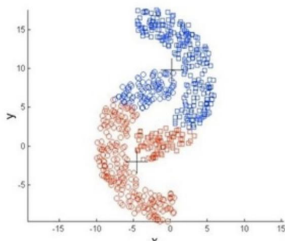
➤ 主要需要调节的参数仅是簇数 k 。

k-均值聚类的缺点

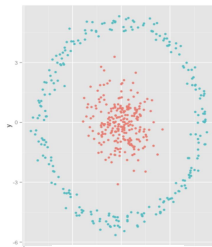
- 簇数 k 的值很难确定。
- 仅适用于数值型数据，对分类型数据不直接适用，因为对于分类型数据，均值向量没有定义。
- 对于非凸/非球形的簇，聚类效果不佳。



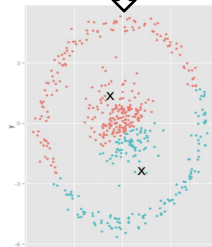
Original Points



K-means



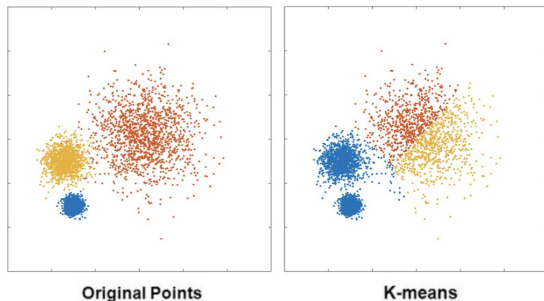
Original Points



K-means

k-均值聚类的缺点

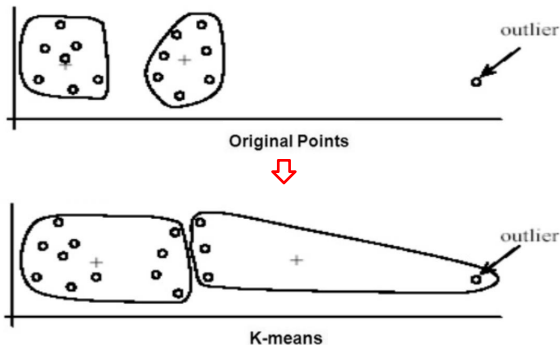
- 对于各簇中样本规模差异太大、不同密度的簇，聚类效果不佳。



- 最终的聚类结果和初始簇中心的选择有关，容易陷入局部最优。

k-均值聚类的缺点

- 对噪声/离群点很敏感：离群点是离其它数据点很远的
的数据点，离群点可能是由于数据记录的错误造成的。



因此，在利用 k -均值算法执行聚类分析前，要去除离群点。

k-中心点聚类算法

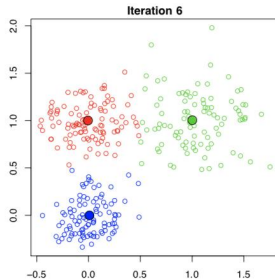
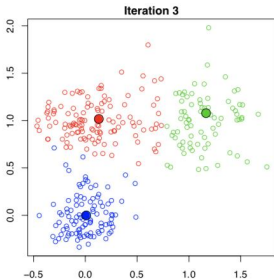
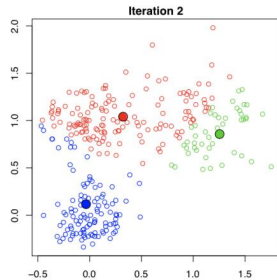
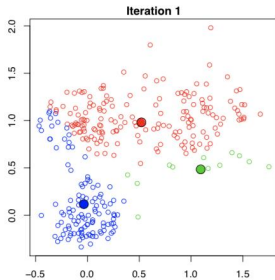
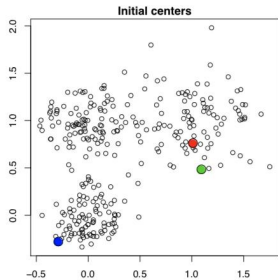
➤ 算法流程:

- ✳️ 步骤1: 随机选择 k 个样本作为初始的簇中心;
- ✳️ 步骤2: 对剩余的每个样本, 将其划分到距离它最近的簇中心所在的簇中, 从而形成 k 个簇;
- ✳️ 步骤3: 重新计算簇中心, 即对于每个簇 C_k , 找到该簇的一个样本点, 称为medoid, 使得该簇中所有其它点到该样本点的距离之和最小, 即

$$\underset{x \in C_k}{\operatorname{argmin}} \sum_{x' \in C_k} \|x - x'\|^2$$

- ✳️ 步骤4: 如果簇中心没有任何改变, 算法停止, 否则回到步骤2

k-中心点聚类流程

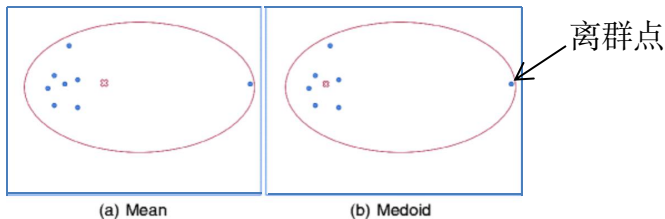


k-均值聚类与k-中心点聚类对比

- k -中心点算法与 k -均值算法比较：
 - ✳ 聚类过程类似，区别在于簇中心点的选取；
 - ✳ 每次迭代都会降低误差平方和准则函数的值；
 - ✳ 最终的聚类结果都受初始簇中心选择的影响；
 - ✳ 算法都会收敛，但都容易陷入局部最优；
 - ✳ 当样本集中存在噪声/离群点时， k -中心点算法比 k -均值算法更鲁棒；
 - ✳ k -中心点算法每次迭代中计算簇的medoid的代价是 $O(n^2)$ ，而 k -均值算法计算均值向量的代价为 $O(n)$ 。

k-中心点聚类的特点

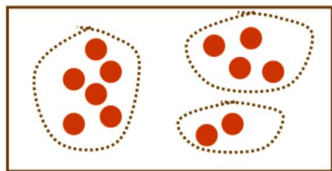
- 对噪声/离群点更鲁棒：因为簇的均值向量易受离群点影响，但簇的medoid可以削弱离群点的影响。



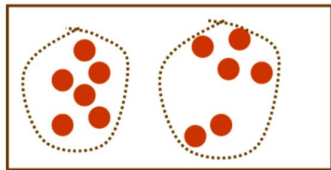
- 算法的时间复杂度更高： $O(tkn^2)$ ，其中， n 是样本数， k 是簇的数目， t 是算法迭代次数。因为每次迭代中计算簇的medoid的代价比计算簇均值向量的代价大的多。

基于层次的聚类算法

- 基于划分的聚类方法将样本集划分为不同的簇，不能反应簇与簇之间的联系，例如包含与被包含的关系。
- 层次聚类旨在构建具有层次结构的簇。

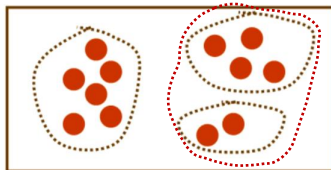


?



聚成三类还是两类呢？

层次聚类：一个大簇中包含若干个小簇，小簇中又可以包含更小的簇。

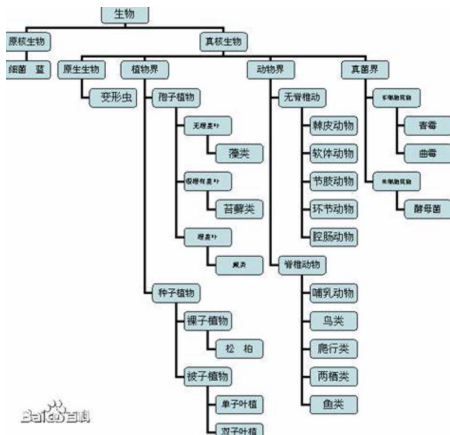


层次聚类算法的应用

- 层次分类法广泛用于很多领域：生物学、图书管理学、网址管理等。

分类：

- 自然科学总论(667)
- 文化、科学、教育、体育(502)
- 政治、法律(331)
- 数理科学与化学(292)
- 哲学、宗教(259)
- 经济(190)
- 工业技术(139)
- 综合性图书(108)
- 历史、地理(99)
- 社会科学总论(77)
- 生物科学(75)
- 医药、卫生(65)
- 文学(63)
- 农业科学(55)
- 语言、文字(36)
- 天文学、地球科学(34)
- 环境科学、安全科学(20)
- 艺术(15)

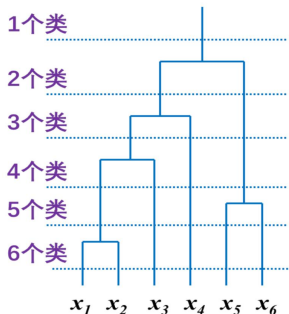


360百科

层次聚类的两种方法

➤ 合并式聚类 (agglomerative clustering)

- ✿ 自底向上，合并
- ✿ 初始每个样本自成一簇，然后递归地合并最相似的两个簇，直到簇的数目为预定义的 c 个
- ✿ 需要定义簇与簇之间的相似度

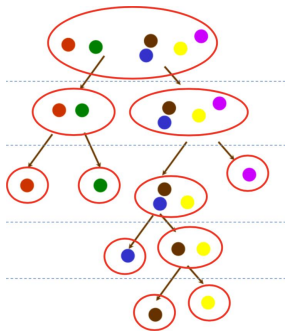


如果两个样本在前期被合并到一个簇，以后它们一直属于同一个簇。

层次聚类的两种方法

➤ 分裂式聚类 (divisive clustering)

- ❁ 自顶向下，分裂
- ❁ 初始所有样本在一个簇中，即根簇，然后递归地将根簇分裂成若干个子簇，直到簇的数目为预定义的 c 个



每次分裂时，需要考虑所有可能的分裂组合，代价比较大。

如果两个样本在前期被分到两个不同的簇，以后它们一直属于不同的簇。

簇与簇之间的相似性度量

➤ 常用的4种方法度量簇与簇之间的相似性:

✧ 最小距离: 两个簇中相距最近的两个样本之间的距离

$$dist_{min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{x \in \mathcal{D}_i, x' \in \mathcal{D}_j} \|x - x'\|$$

✧ 最大距离: 两个簇中相距最远的两个样本之间的距离

$$dist_{max}(\mathcal{D}_i, \mathcal{D}_j) = \max_{x \in \mathcal{D}_i, x' \in \mathcal{D}_j} \|x - x'\|$$

✧ 平均距离: 两个簇中两两样本之间的平均距离

$$dist_{avg}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{n_i n_j} \sum_{x \in \mathcal{D}_i} \sum_{x' \in \mathcal{D}_j} \|x - x'\|$$

✧ 均值距离: 两个簇的均值向量之间的距离

$$dist_{mean}(\mathcal{D}_i, \mathcal{D}_j) = \|\mu_i - \mu_j\|$$

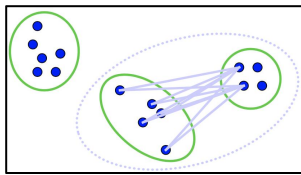
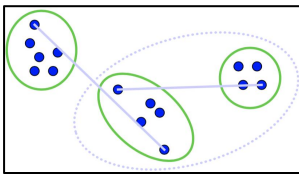
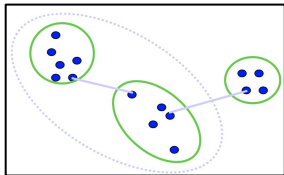
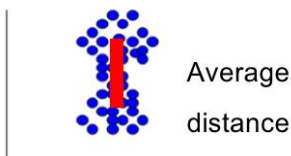
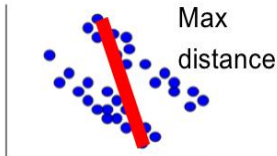
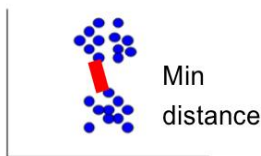
层次聚类算法(了解)

➤ 当**合并式**层次聚类算法采用

✧ **最小距离**时，称为最近邻聚类，也称为**单连接**聚类。

✧ **最大距离**时，称为最远邻聚类，也称为**全连接**聚类。

✧ **平均距离**时，称为**平均连接**聚类。

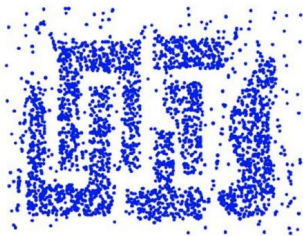


层次聚类算法的特点

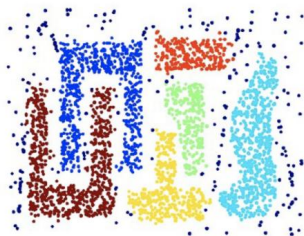
- 可以发现簇的层次结构。
- 无论是合并还是分裂算法，都是局部最优的算法，因为合并/分裂的操作不能撤销，每一步合并或分裂时求得的是当前一步的最优聚类，但是前期被合并的样本，以后无法再分开，前期被分开的样本，以后就无法再合并，导致不能保证求得全局最优解。
- 时间复杂度至少是 $O(t * n^2)$ ， t 为迭代次数， n 为样本总数，因此算法拓展性不好，不适合大数据集。

基于密度的聚类算法

- 基于划分和基于层次的聚类算法只能发现球状簇，而基于密度的聚类算法能发现任意形状的簇，且对噪声/离群点不敏感。
- 基于密度的聚类的思想是：把簇看作是特征空间中由低密度样本区域分隔开的高密度样本区域。



原始数据



数据中的簇

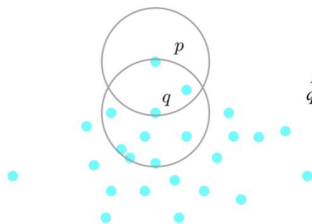
DBSCAN算法

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): 能在具有噪声的数据集中发现任意形状的簇, 它将簇定义为**密度相连的点的最大集合**。
- 算法的两个输入参数: 邻域半径 ϵ 和整数 $minPts$
- 基本概念:
 - ✧ 点 x 的邻域 $N_\epsilon(x)$: 以 x 为球心, 半径为 ϵ 的球内区域。
 - ✧ 核心点: 如果 $N_\epsilon(x)$ 内包含的样本点个数 $\geq minPts$, 则称 x 为核心点。
 - ✧ 边界点: 如果 $N_\epsilon(x)$ 内包含的样本点个数 $< minPts$, 且 $N_\epsilon(x)$ 内包含核心点, 则称 x 为边界点。
 - ✧ 噪声点: 除了核心点和边界点以外的点。

DBSCAN算法

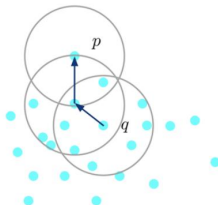
➤ 关键概念:

- ❁ 密度直达: 如果点 q 是核心点, 且在 $N_\epsilon(q)$ 内包含一个点 p , 则称点 p 是从点 q 密度直达的。
- ❁ 密度可达: 如果存在一个样本链 p_1, p_2, \dots, p_n , 使得 p_{i+1} 是从 p_i 密度直达的, $\forall i = 1, 2, \dots, n-1$, 则称点 p_n 是从点 p_1 密度可达的。



p 可由 q 密度直达
 q 不可由 p 密度直达

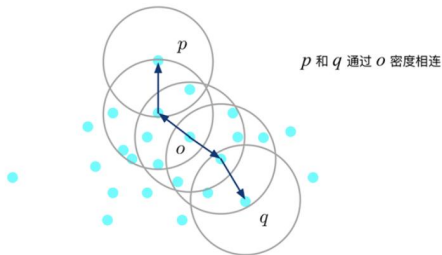
p 可由 q 密度可达
 q 不可由 p 密度可达



DBSCAN算法

➤ 关键概念:

- ❁ **密度相连**: 如果存在一个点 o , 使得点 p 和点 q 都是从点 o 密度可达的, 则称点 p 和点 q 是密度相连的。



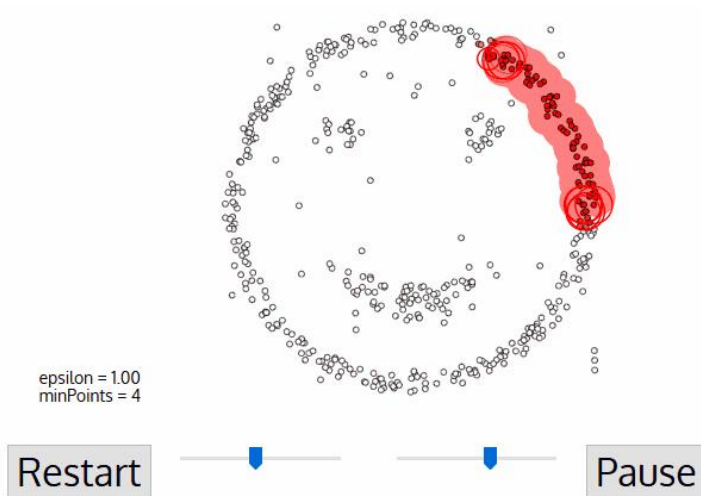
➤ DBSCAN算法:

“簇” 定义为: **密度相连**的样本的最大集合。

DBSCAN算法

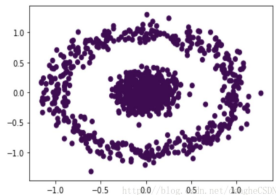
- DBSCAN算法基本描述：由核心点出发，找到由该核心点密度可达的所有样本形成“簇”。
- DBSCAN算法的流程为：
 - 根据给定的邻域参数 ϵ 和MinPts确定所有的核心点；
 - 对每一个核心点
 - 如果该核心点未被加入某个簇，找到由其密度可达的所有样本生成聚类的一个“簇”
 - 直到所有的核心点都被加入一个簇

DBSCAN聚类过程

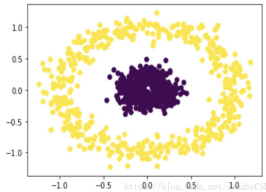


DBSCAN算法

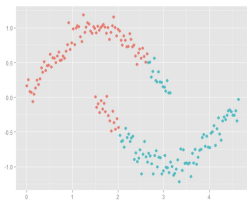
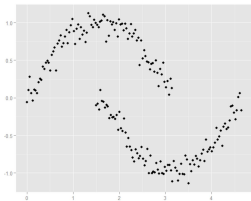
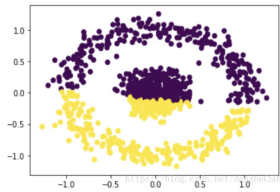
原始数据



DBSCAN




K-means



DBSCAN算法的优缺点

- 对噪声不敏感：该算法能够较好地判断噪声/离群点，并且即使错判，对最终的聚类结果也没什么影响。
- 能发现任意形状的簇：DBSCAN 靠不断连接邻域内高密度点来发现簇，只需要定义邻域半径 ϵ 和MinPts，因此可以发现不同形状、不同大小的簇。
- 缺点：
 - ① 对两个参数的设置敏感，即邻域半径 ϵ 和MinPts。
 - ② 当簇的稀疏程度不同，即簇的密度不均匀时，很难使用该算法。

小结


- 有监督学习依赖于已经标注好类别标签的样本构成的训练集，而无监督学习中训练集中样本的类别标记是未知的，目的是发现样本集中内在的结构或规律。
- 聚类就是把样本集中的样本按照相似的程度划分成不同的簇，使得同一个簇内样本相似程度远大于不同簇的样本的相似程度。好的聚类结果一定簇内样本的相似度高，簇间样本的相似度低。
- 聚类结果受聚类准则、相似性度量、各个特征的量纲标尺的影响。
- 聚类的基本功能就是去挖掘数据中隐藏的知识和结构，解释样本间的内在联系，聚类还可以用于对数据集进行整理、对样本进行自动类别标注、对样本集简化等。

小结

- 聚类的基本流程是：选定聚类所用的特性、选择相似性度量、选择聚类算法，评估聚类结果。
- 聚类常用准则函数：误差平方和准则、最小方差准则和散布准则。
- 基于划分的聚类方法旨在将样本集划分为 k 个互不相交的集合，典型算法有： k -均值 (k -means)和 k -中心点聚类 (k -medoids)，理解两个聚类算法的原理、优点和限制，能够编程实现这两种方法。
- 层次聚类能够构建具有层次结构的簇，构建层次聚类的方法有合并式聚类和分裂式聚类。两种方法都容易陷入局部最优，因为聚类前期的合并/分裂操作是不能撤销的。

重点

小结

- 基于密度的聚类思想是：把簇看作是特征空间中由低密度样本区域分隔开的高密度样本区域，目的是找到位于高密度区域的样本，形成簇。
- DBSCAN算法是一种基于密度的聚类算法，该算法将由一个核心点出发所有密度可达的点构成的集合聚成一个簇。当数据集中各簇的密度不均匀时，DBSCAN算法不适用，因为这种情况下，很难设置全局最优的邻域半径 ϵ 和 MinPts。
- 基于密度的聚类算法能发现任意形状的簇，且对噪声/离群点不敏感，而基于划分和基于层次的聚类算法只能发现球状簇。