媒体数据分析
聚类分析

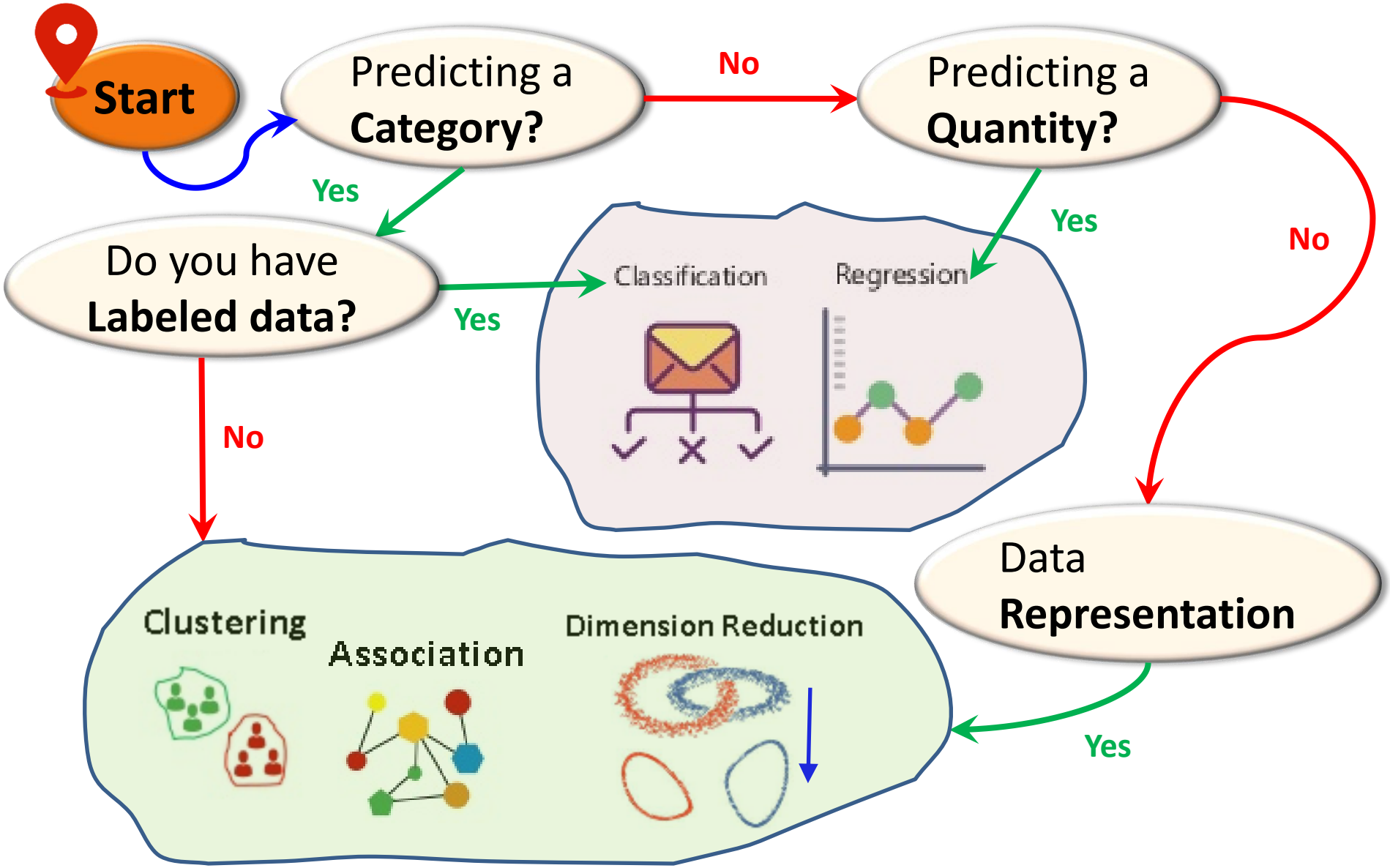
讲师：孙煦雪

浙江傳媒學院
COMMUNICATION
UNIVERSITY
OF ZHEJIANG

课程回顾

- 线性回归
  - 一元线性回归
  - 多元线性回归
  - 正则化
- 非线性回归
- 逻辑回归

知识地图


Start
Predicting a Category?
No
Predicting a Quantity?
Yes
Yes
No
Do you have Labeled data?
Yes
Classification
Regression
No
Data Representation
Clustering
Association
Dimension Reduction
Yes

内容提要

- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

聚类分析概述

- 簇(Cluster)：一个数据对象的集合。

物以类聚，人以群分，这是聚类分析的基本常识。

- 聚类（Clustering）是将对象集合中的对象分类到不同的类或者簇这样的一个过程，使得同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。

聚类分析概述

- 聚类分析的目标

簇内相似，簇间差别大

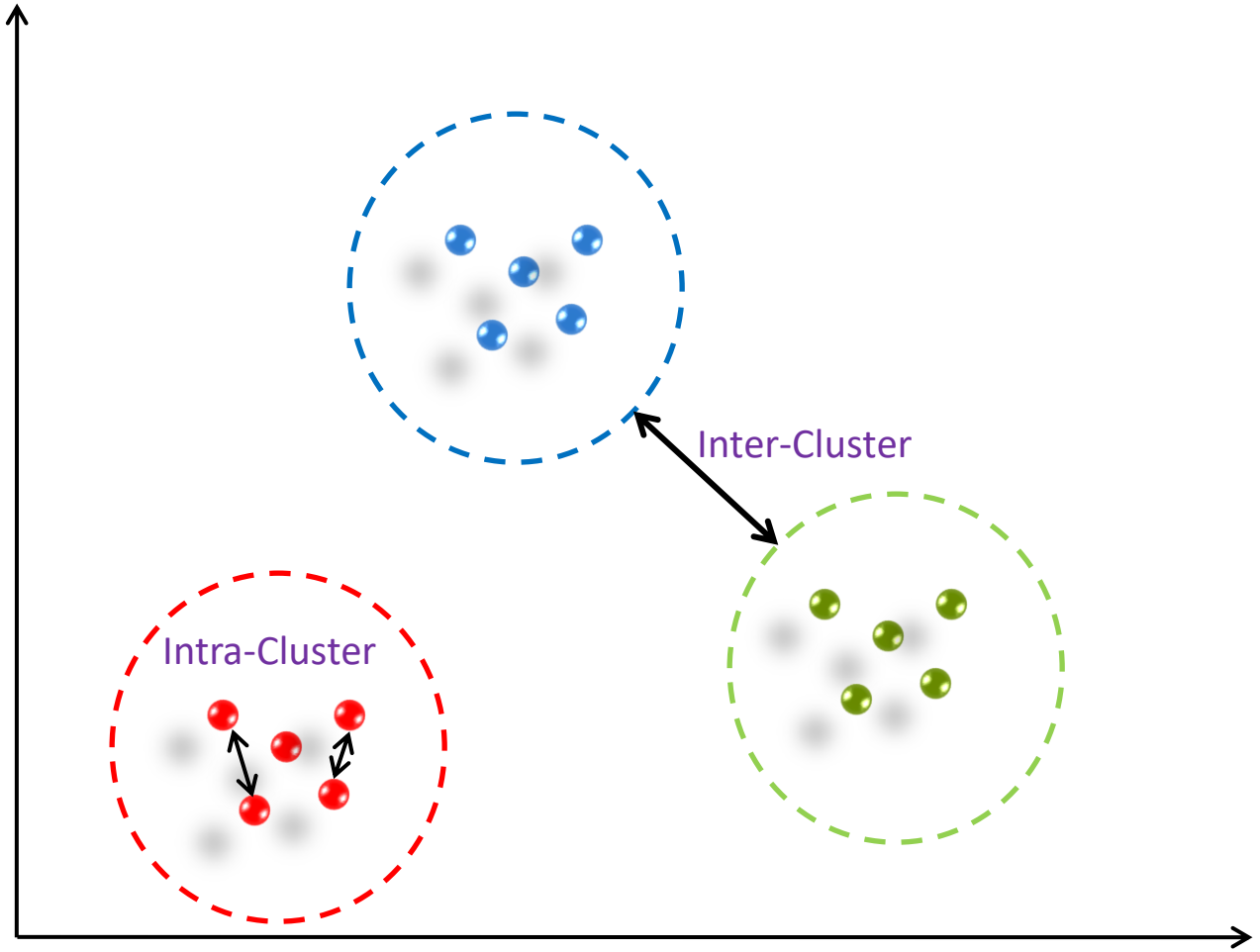
  - 一个簇内的数据尽量相似
  - 不同簇的数据尽量不相似

- 聚类结果的好坏

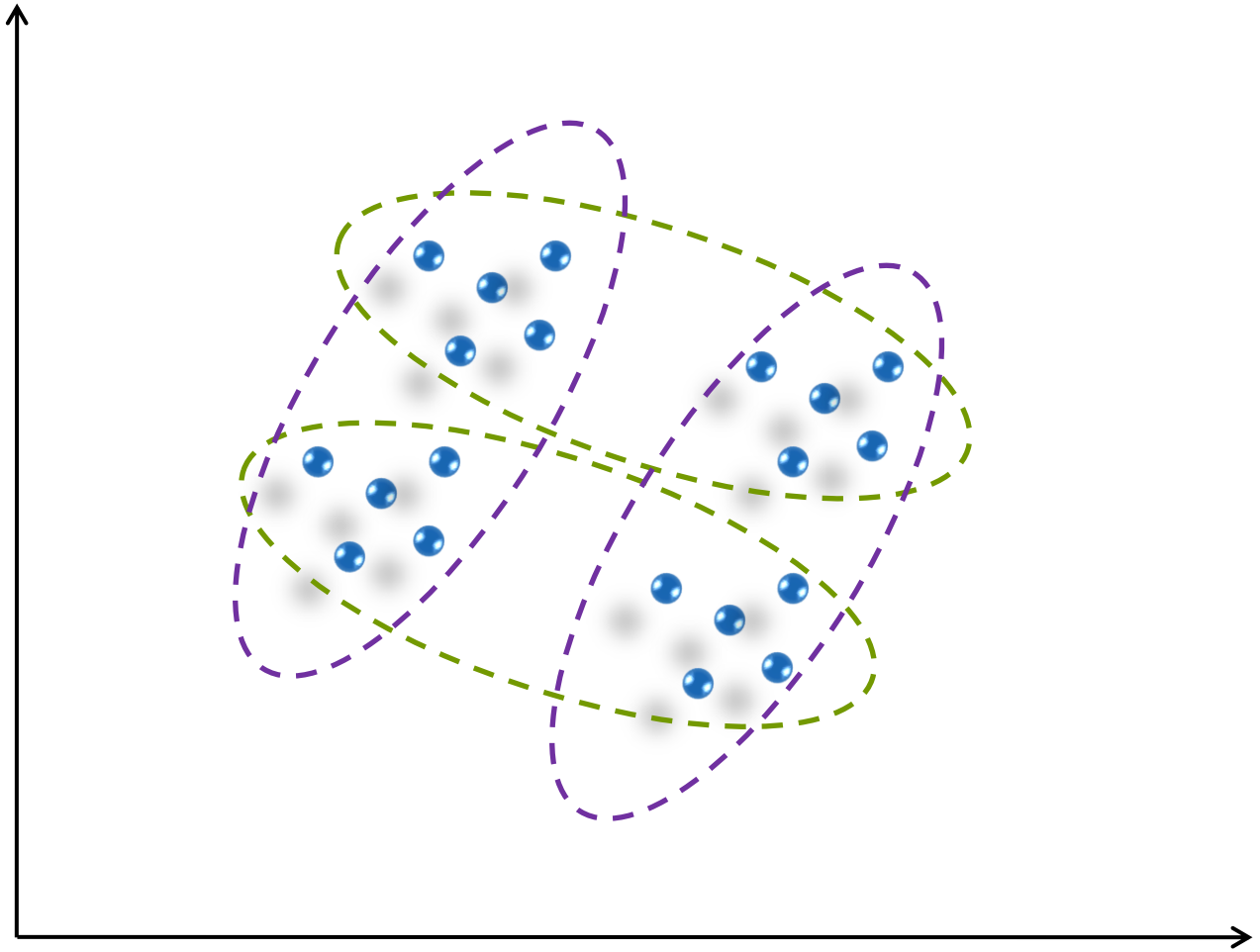
衡量一个聚类分析算法质量，依靠：

  - 相似度测量机制（Similarity measure）是否合适
  - 是否能发现数据背后潜在的、手工难以发现的类知识

聚类分析概述



聚类分析概述



聚类典型的应用

- 数据化运营和个性化推荐

基于在线行为数据（播放、点赞、分享、收藏等），对在线用户（音乐平台或电影平台）进行聚类，以此进行个性化推荐。

- 搜索引擎

对返回的结果进行聚类，使用户迅速定位到所需要的信息。

- 新媒体文章价值评估和运营决策

从粉丝、传播力度和文章信息多个角度，对历史发布在新媒体的文章（例如微信公众号文章）进行聚类，为今后文章的内容和采编方向做指导。

- 基于文档信息的推荐

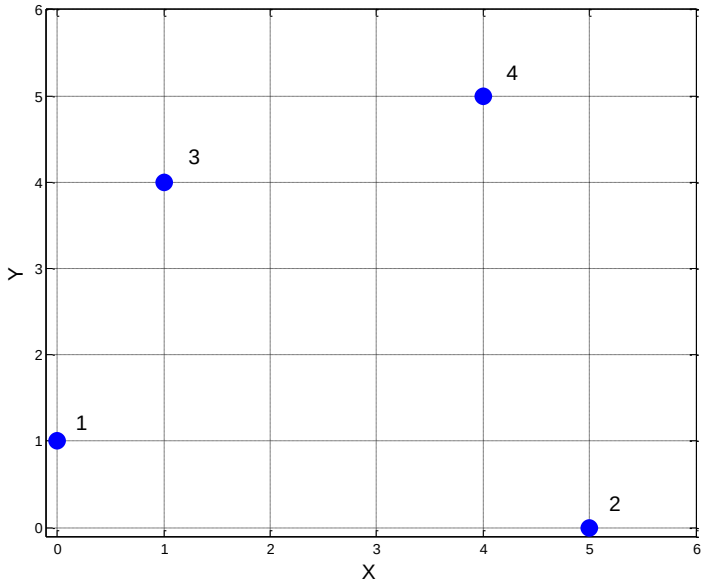
对用户感兴趣的文档（例如用户浏览过的网页）聚类，从而发现用户的兴趣模式并用于信息过滤和信息主动推荐等服务。

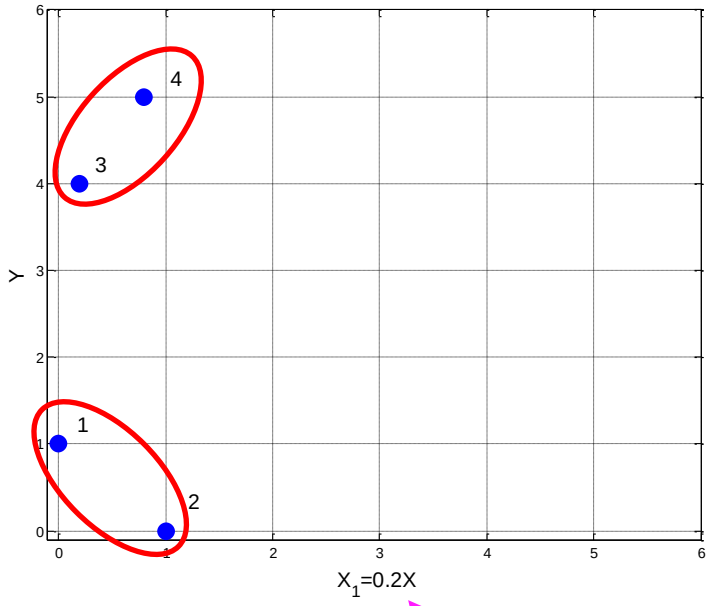
数据挖掘对聚类的典型要求

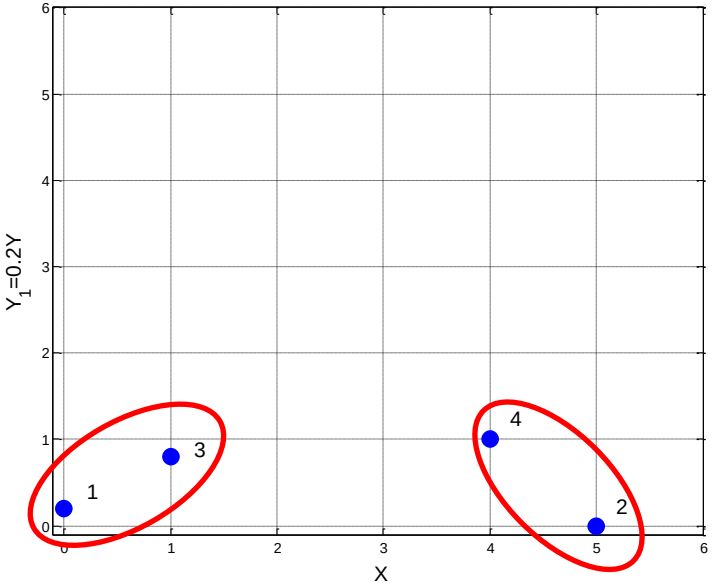
- 可伸缩性
- 能够处理不同类型的属性
- 能发现任意形状的簇
- 在决定输入参数的时候，尽量不需要特定的领域知识
- 能够处理噪声和异常
- 对输入数据对象的顺序不敏感
- 能处理高维数据
- 能产生一个好的、能满足用户指定约束的聚类结果
- 结果是可解释的、可理解的和可用的

https://www.bilibili.com/video/BV1qr4y1P7p9/ ~ 1’45

聚类分析中的实际考虑

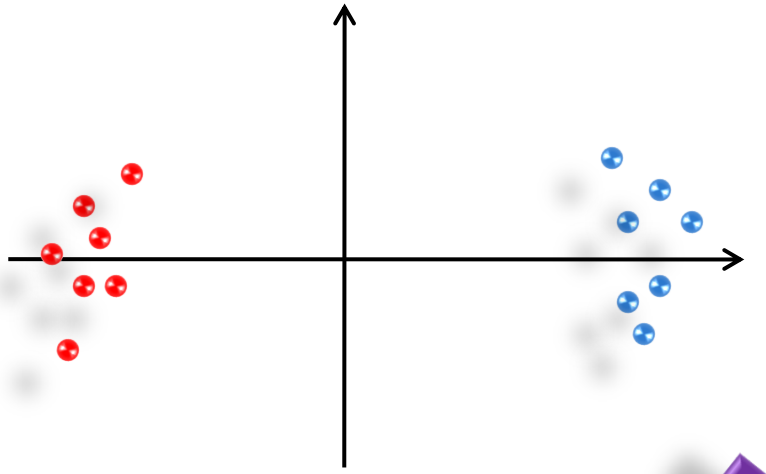


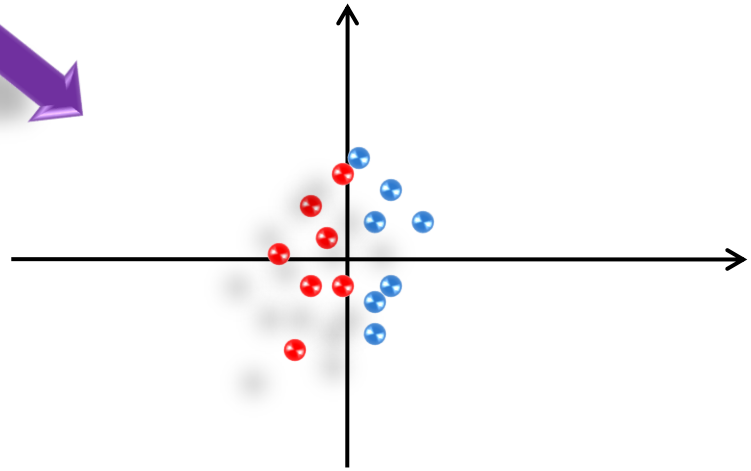




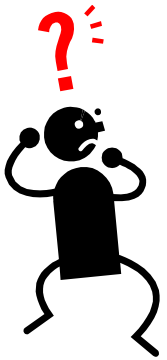
Scaling matters!

聚类分析中的实际考虑





Normalization or Not



内容提要

- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

聚类方法分类

按照聚类方法的主要思路的不同，聚类方法分为：

- 划分聚类

  基于一定标准构建数据的划分，例如k-means等

- 层次聚类

  对给定数据对象集合进行层次的分解，例如BIRCH算法

- 基于密度的聚类

  基于数据对象的相连密度评价，例如DBSCAN算法

- 基于网格的聚类

  将数据空间划分成为有限个单元（Cell）的网格结构，基于网格结构进行聚类，例如STING算法

- 基于模型的聚类

  给每一个簇假定一个模型，然后去寻找能够很好的满足这个模型的数据集，例如高斯混合模型

内容提要

- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

划分聚类

对于给定的数据集，划分聚类方法首先创建一个初试划分，然后采用一种迭代的重定位技术，尝试通过对象在划分间的移动来改进划分，直到使评价聚类性能的评价函数的值达到最优为止。

划分聚类方法以距离作为数据集中不同数据间的相似性度量，将数据集划分成多个簇。包括k均值（k-means），k中心点等方法。

常用的距离函数：

- 明可夫斯基距离（Minkowski） $D(x,y)=\left(\sum_{u=1}^{n}|x_u-y_u|^p\right)^{\frac{1}{p}}$
- 欧几里得距离： p=2
- 曼哈顿距离: p=1
- 切比雪夫距离（Chebyshev): p->∞

  二维向量: max(|x1−x2|,|y1−y2|)

划分聚类主要思想

给定一个包含n个数据对象的数据集，划分聚类方法将数据对象的数据集进行k个划分，每个划分表示一个簇(类)，并且k≤n，同时满足两个条件：

- 每个簇至少包含一个对象
- 每个对象属于且仅属于一个簇

对于给定k，划分聚类方法首先给出一个初始的划分，然后采用一种迭代的重定位技术（**反复迭代**），尝试通过对象在划分间移动来改进划分，使得每一次改进之后的划分方案都较前一次更好。

划分聚类的评价函数

好的划分是指同一簇中的对象之间尽可能“接近”，不同簇中的对象之间尽可能“远离”。

评价函数着重考虑两方面，即每个簇中的对象应该是紧凑的，各个簇间的对象的距离应该尽可能远。

实现这种考虑的一种直接方法就是观察聚类C的类内差异w(C)和类间差异b(C)。

类内差异衡量类内的对象之间的紧凑性，类间差异衡量不同类之间的距离。

划分聚类的评价函数

- 类内差异可以用距离函数来表示，最简单的就是计算类内的每个对象点到它所属类的中心的距离的平方和。
- 类间差异定义为类中心之间距离的平方和。

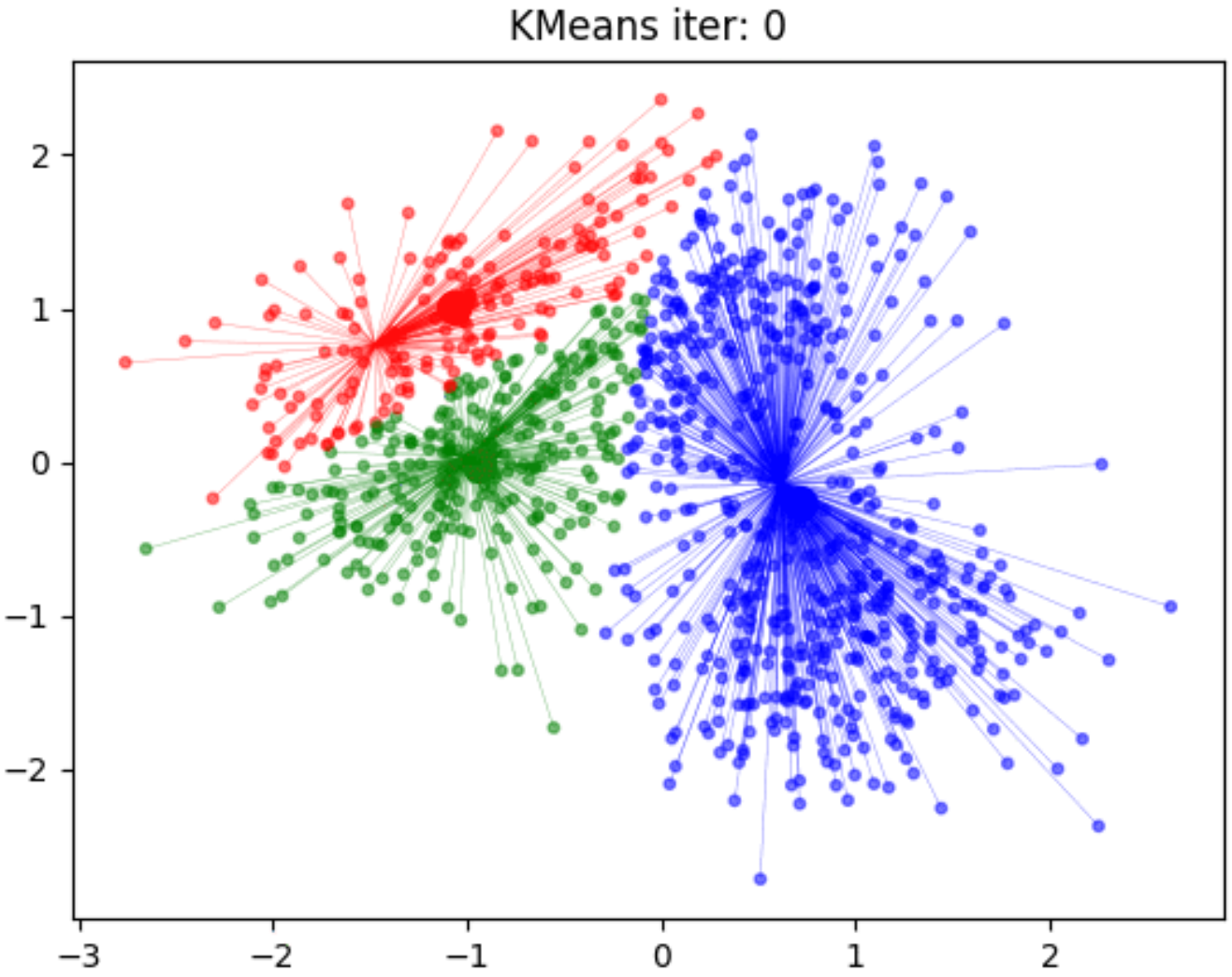
类间距离的度量还包括：

最短距离法：定义两个类中最靠近的两个元素间的距离为类间距离。

最长距离法：定义两个类中最远的两个元素间的距离为类间距离。

类平均法：它计算两个类中任意两个元素间的距离的平均值作为类间距离。

k均值聚类



k均值聚类原理

Step1：随机初始化K个聚类中心，即K个类中心向量

Step2：对每个样本，计算其与各个类中心向量的距离，并将该样本指派给距离最小的类

Step3：更新每个类的中心向量，更新的方法为取该类所有样本的特征向量均值。

Step4：直到各个类的中心向量不再发生变化，退出。

k均值算法的目标函数E定义为 $E = \sum_{i=1}^{k} \sum_{x \in C_i} [d(x, \bar{x}_i)]^2$

https://www.bilibili.com/video/BV1mf4y1k7UC/

k均值算法描述

输入：所期望的簇的数目k，包含n个对象的数据集D

输出：k个簇的集合

① 从D中任意选择k个对象作为初始簇中心；

② repeat

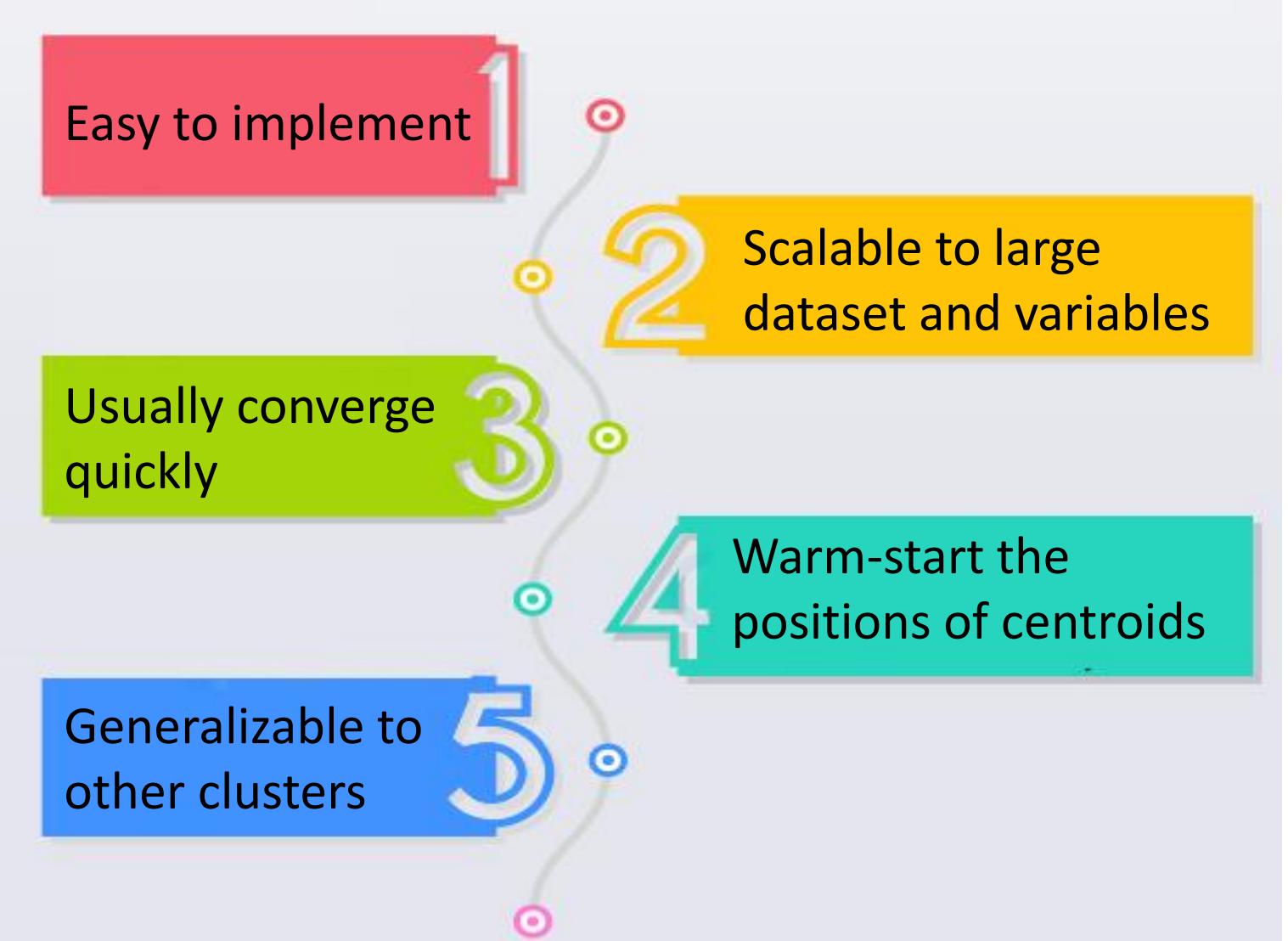
③ 将每个点指派到最近的中心，形成k个簇；

④ 重新计算每个簇的中心；

⑤ 计算目标函数E；

⑥ until 目标函数E不再发生变化或中心不再发生变化；

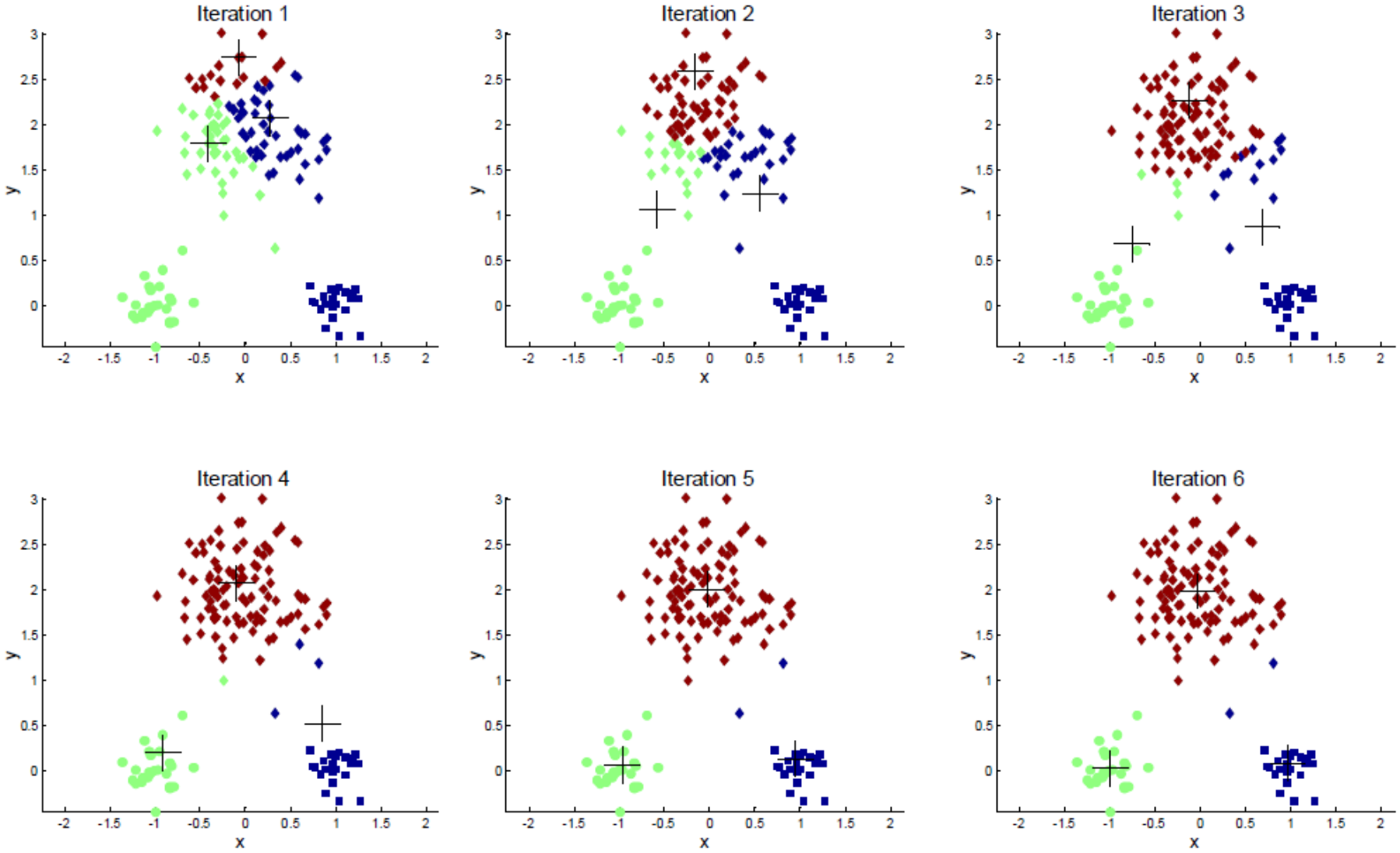
k均值算法主要优点


1
Easy to implement
2
Scalable to large dataset and variables
3
Usually converge quickly
4
Warm-start the positions of centroids
5
Generalizable to other clusters

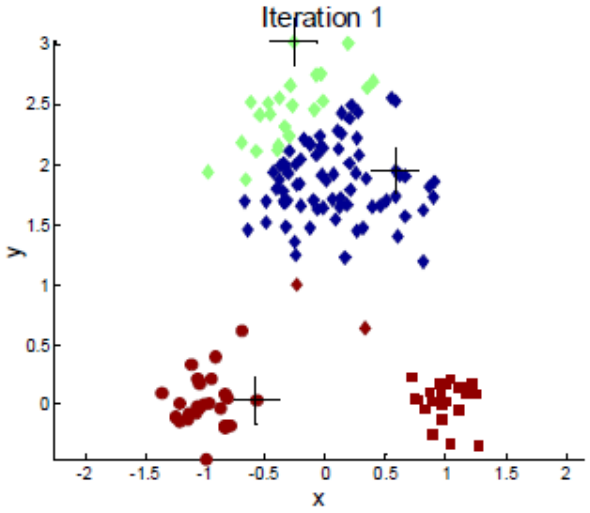
k均值算法主要缺点

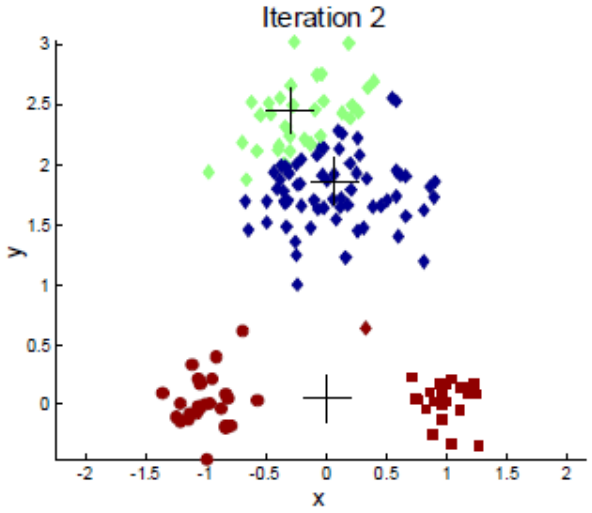
- Manually choose *K*
- Curse of dimensionality
- Converge to local minima
- Dependent of initial values or partitions
- Not applicable for many types of clusters (varying size and densities)
- Sensitive to outliers

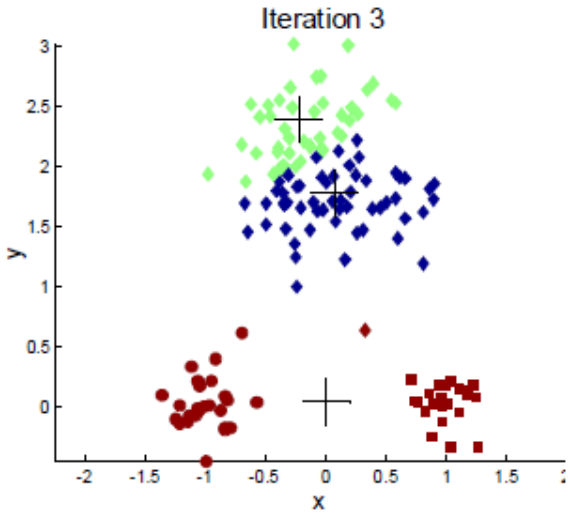
初始聚类中心的影响

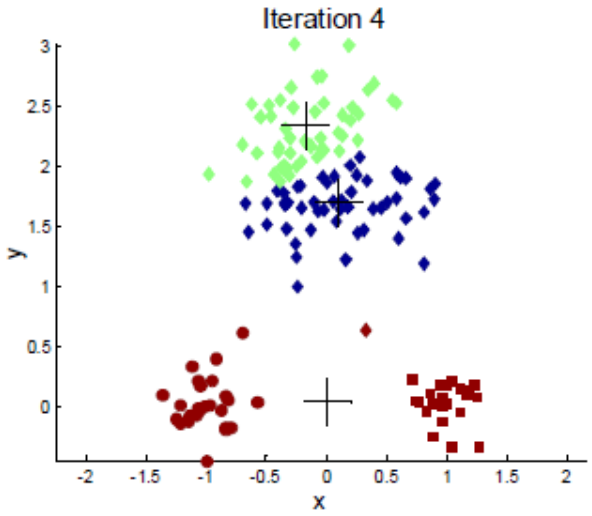

Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y

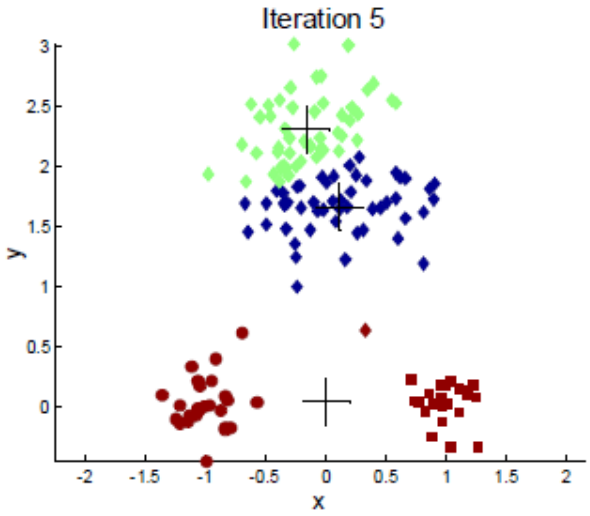
初始聚类中心的影响


Iteration 1
x
y


Iteration 2
x
y


Iteration 3
x
y


Iteration 4
x
y


Iteration 5
x
y

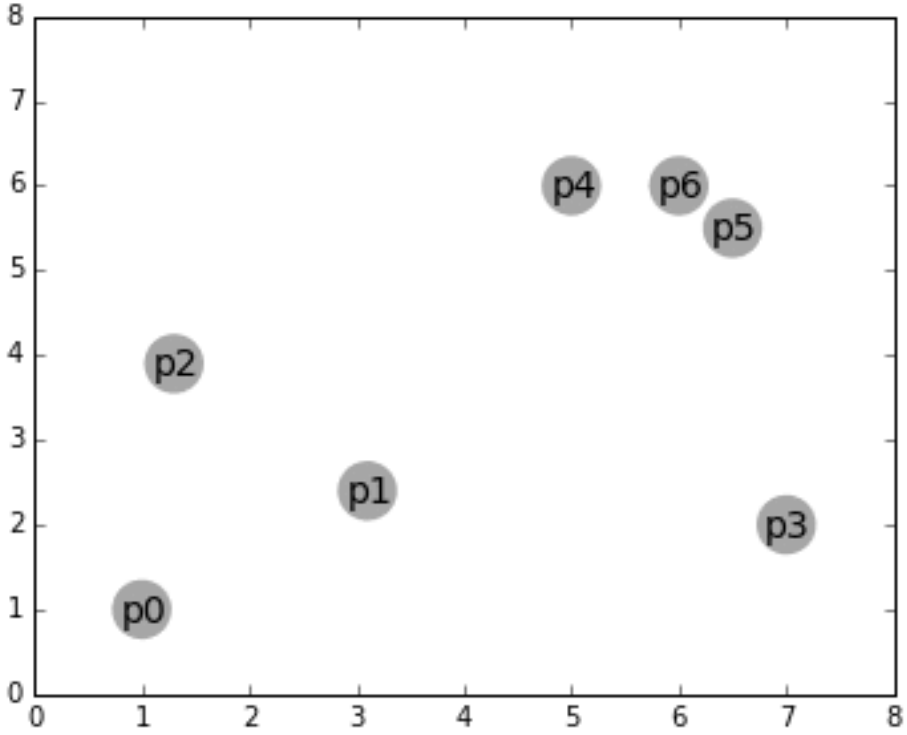
实战分析

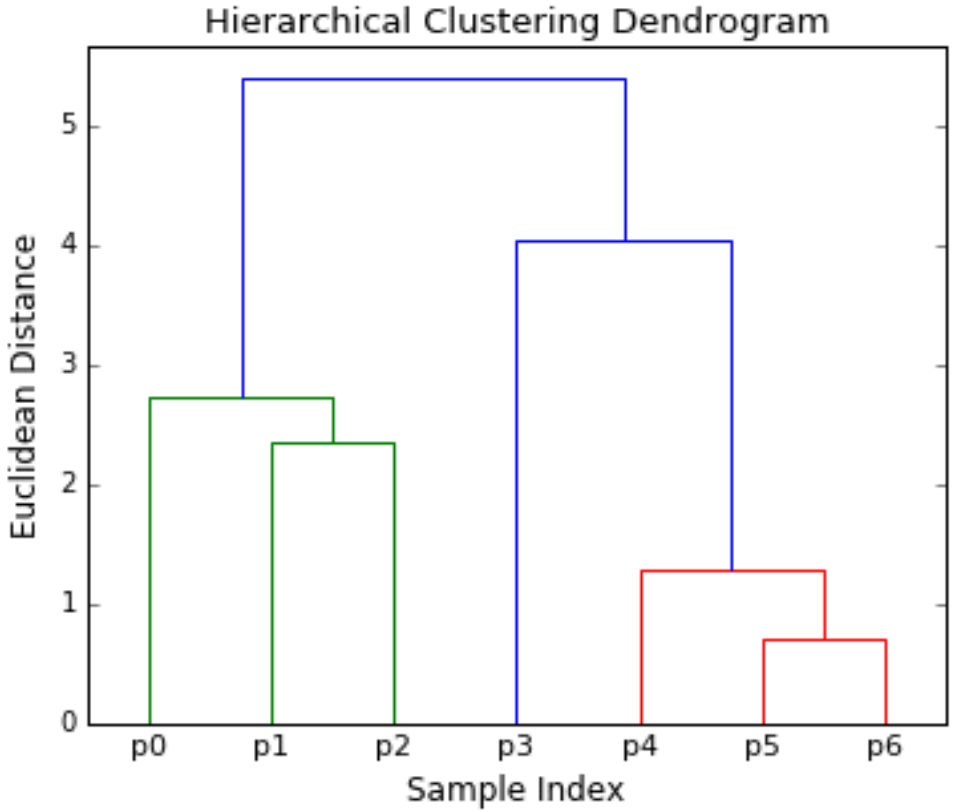
- K均值聚类分析编写
- 鸢尾花数据集K-means聚类
- 农村居民人均可支配收入聚类分析

内容提要

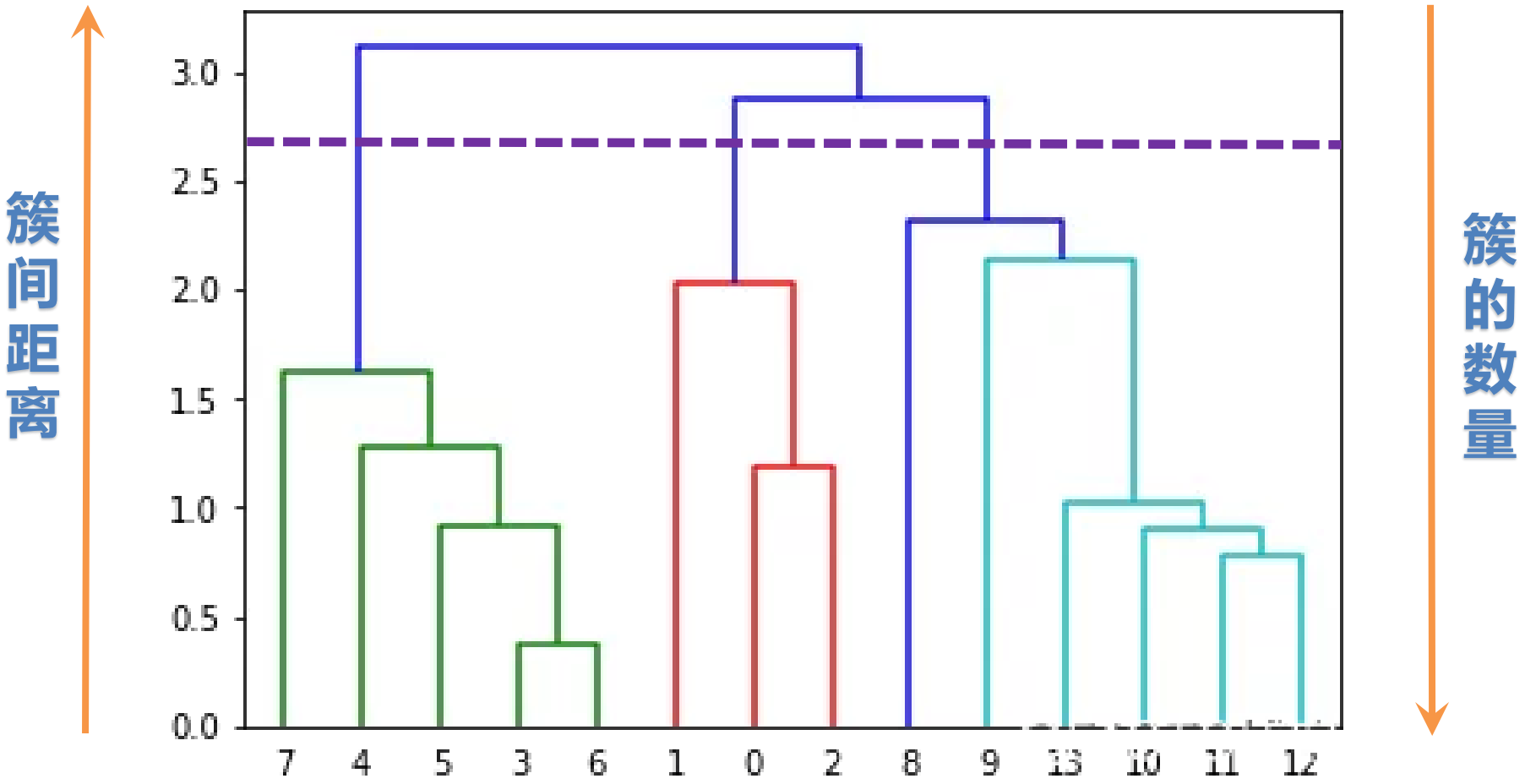
- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

层次聚类原理


p0
p1
p2
p3
p4
p5
p6


Hierarchical Clustering Dendrogram
Euclidean Distance
Sample Index
p0
p1
p2
p3
p4
p5
p6

层次聚类原理



层次聚类原理

通过递归地对数据对象进行合并或者分裂，直到满足某种终止条件为止。

①将n个样本分成n簇，每个样本划为一簇。

②每次将具有最小距离的两个类合并，合并后重新计算簇与簇之间的距离，重复这个步骤，直到所有样本都成一簇。

https://www.bilibili.com/video/BV18r4y1z7vs/

层次聚类

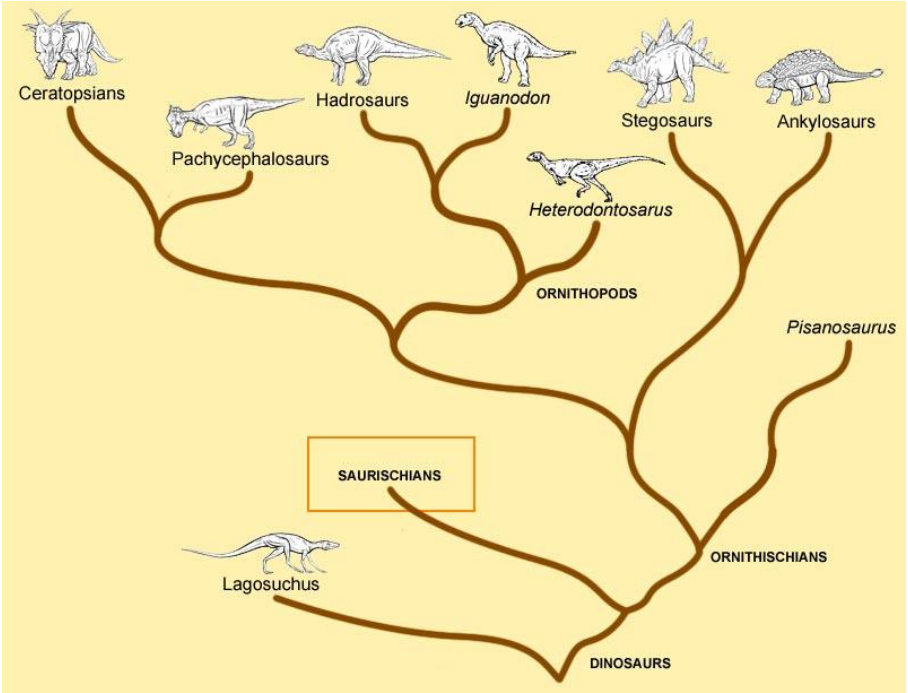
根据层次分解，层次聚类方法分为：

- 凝聚型聚类方法（agglomerative clustering）

  自底向上（合并）

- 分裂型聚类方法（divisive clustering）

  自顶向下（分裂）



层次聚类

- 自底向上的凝聚层次聚类（AGNES算法）

  输入：n个对象，终止条件簇的数目k

  输出：k个簇

  1：将每个对象当成一个初始簇

  2：Repeat

  3：　根据两个簇中最近的数据点找到最近的两个簇

  4：　合并两个簇，生成新的簇的集合

  5：Until达到定义的簇的数目

- 自顶向下的分裂层次聚类（DIANA算法）

  首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达到了某个终止条件，例如达到了某个希望的簇数目，或者两个最近的簇之间的距离超过了某个阈值。

https://www.bilibili.com/video/BV1bA4y197JR/

簇间距离度量方法

1）簇间最小距离 （Single linkage）

是指用两个簇中所有数据点的最近距离代表两个簇的距离。

2）簇间最大距离（Complete linkage）

是指用两个簇所有数据点的最远距离代表两个簇的距离。

3）簇间均值距离

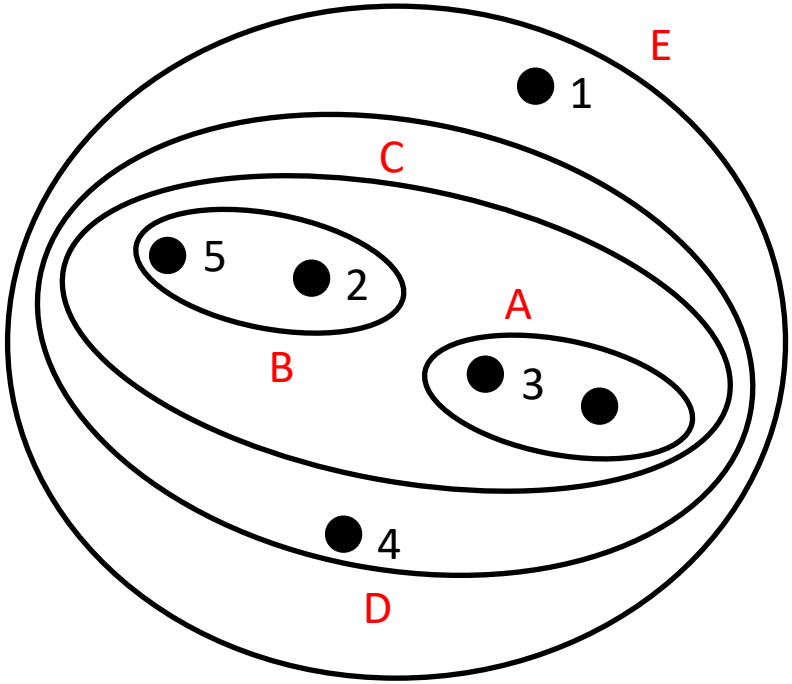
是指用两个簇各自中心点之间的距离代表两个簇的距离。

4）簇间平均距离

用两个簇所有数据点间的距离的平均值代表两个簇的距离。

最小最大度量代表了簇间距离度量的两个极端，它们趋向对离群点或噪声数据过分敏感。使用均值距离和平均距离是对最小和最大距离之间的一种折中方法，而且可以克服离群点敏感性问题。

最小最大距离



MIN （single linkage)

https://www.bilibili.com/video/BV1KJ411U73w?p=3

实战分析

- 层次聚类算法简单例子
- 地区经济数据聚类分析

Hierarchical Clustering算法函数

a）sklearn.cluster.AgglomerativeClustering

b）主要参数(详细参数)

n_clusters：聚类的个数

linkage：指定层次聚类判断相似度的方法，有以下三种：

ward：组间距离等于两类对象之间的最小距离。（即single-linkage聚类）

average：组间距离等于两组对象之间的平均距离。（average-linkage聚类）

complete：组间距离等于两组对象之间的最大距离。（complete-linkage聚类）

c）主要属性

labels_： 每个数据的分类标签

d）算法示例：代码中有详细讲解内容

内容提要

- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

密度聚类原理

基于密度的聚类方法以数据集在空间分布上的稠密程度为依据进行聚类，无需预先设定簇的数量，特别适合对于未知内容的数据集进行聚类。

基于密度聚类方法的基本思想是：只要一个区域中的点的密度大于某个阈值，就把它加到与之相近的聚类中去，对于簇中每个对象，在给定的半径的 ε 邻域中至少要包含最小数目（MinPts）个对象。

基于密度的聚类方法的代表算法为DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类）算法。

DBSCAN聚类算法相关基本术语

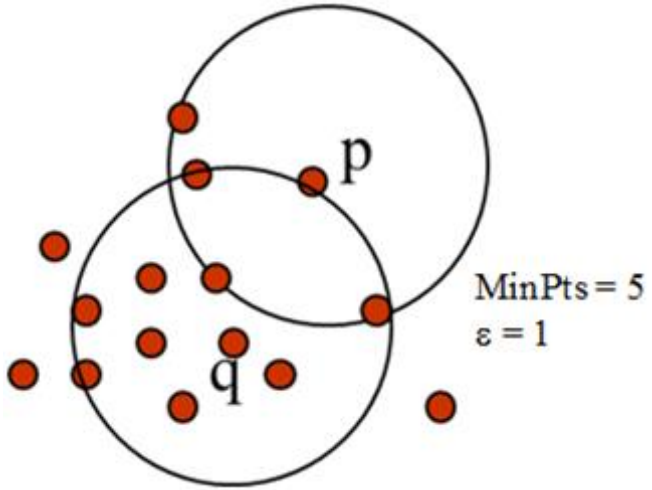
对象的 ε 邻域：给定对象半径为ε内的区域称为该对象的 ε 邻域。

MinPts: 数据对象的 ε 邻域中至少包含的对象数目。

核心对象：如果给定对象ε邻域内的样本点数大于等于MinPts，则称该对象为核心对象。如下图中，q是一个核心对象。

- 直接密度可达

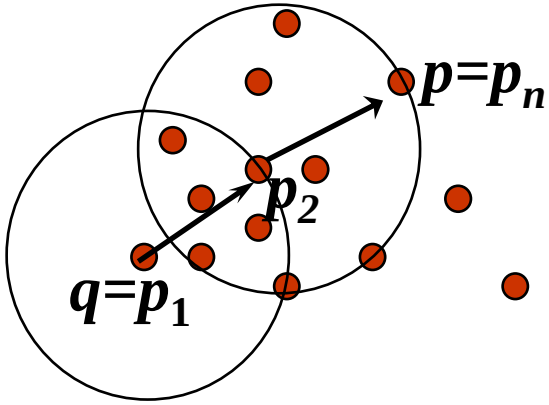
如果 p 在 q 的 ε 邻域内，而 q 是一个核心对象，则称对象 p 从对象 q 出发是直接密度可达的。



DBSCAN聚类算法相关基本术语

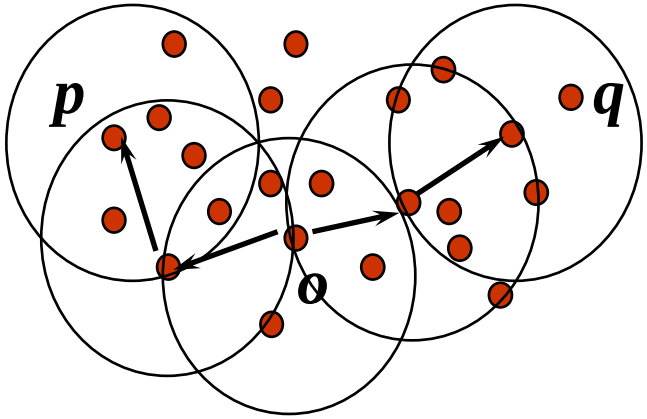
密度可达

如果存在一个对象链 $p_1, \ldots, p_n$，$q=p_1$，$p=p_n$，使得 p_{i+1} 是从 p_i 关于 ε 和 *MinPts* 是直接密度可达的，则对象 *p* 是从对象 *q* 关于 ε 和 *MinPts* 密度可达的

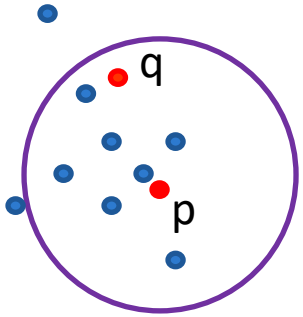


密度相连

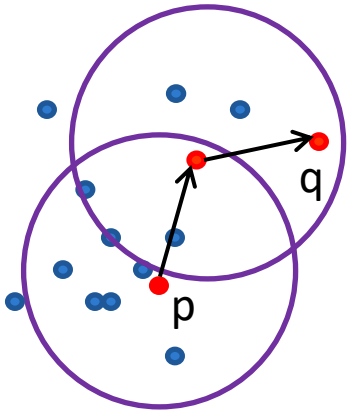
如果存在对象 $o \in D$，使对象 *p* 和 *q* 都是从 *o* 关于 ε 和 *MinPts* 密度可达的，那么对象 *p* 和 *q* 是关于 ε 和 *MinPts* 密度相连的



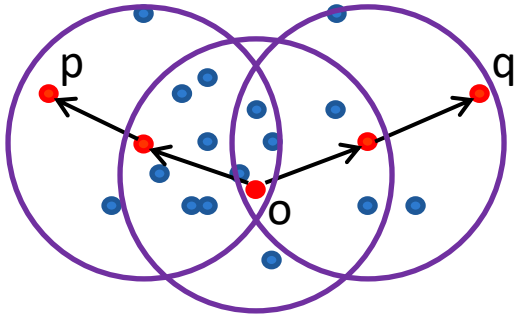
DBSCAN聚类算法相关基本术语



directly density reachable

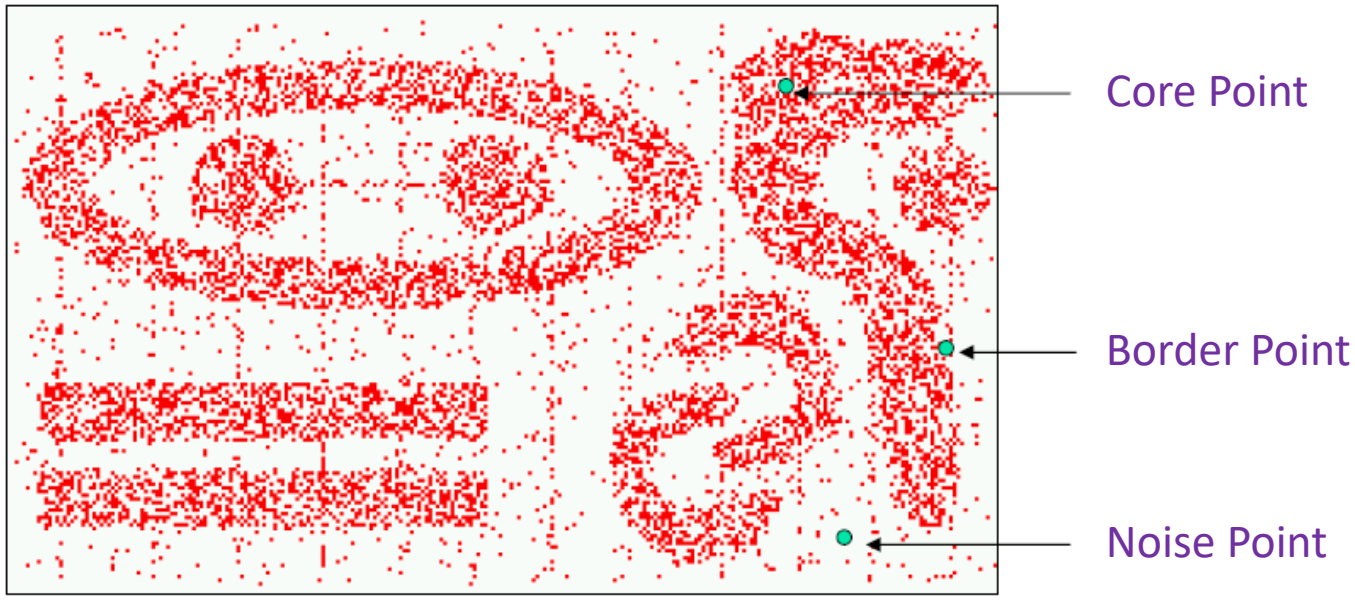


density reachable



density connected

DBSCAN聚类算法



Core Point: points with high density

Border Point: points with low density but in the neighbourhood of a core point

Noise Point: neither a core point nor a border point

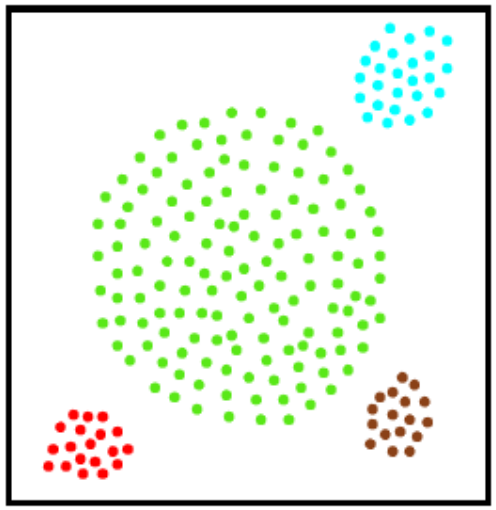
DBSCAN聚类算法

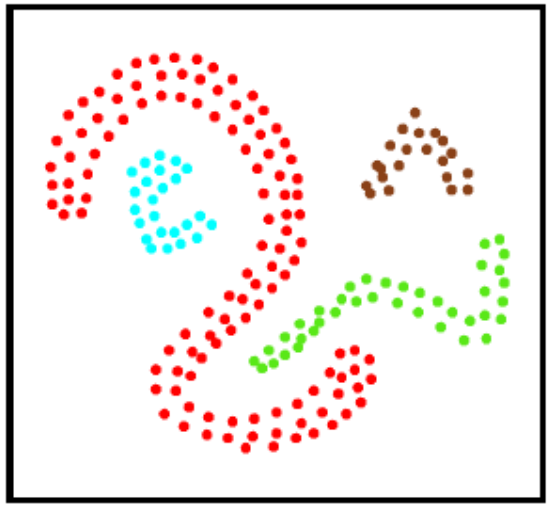
A cluster is defined as the maximal set of density connected points.

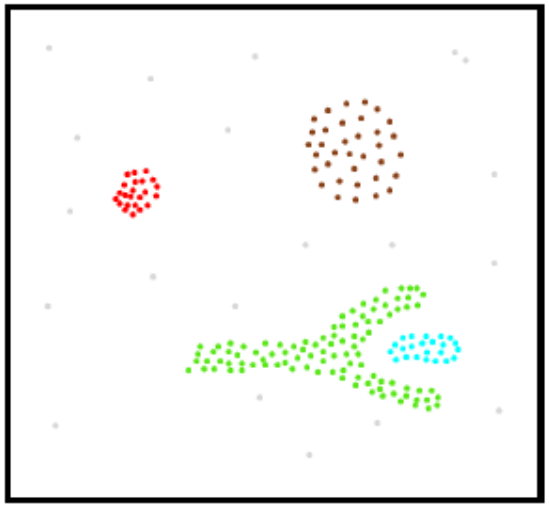
Start from a randomly selected unseen point P.

If P is a core point, build a cluster by gradually adding all points that are density reachable to the current point set.

Noise points are discarded (unlabelled).







DBSCAN聚类算法

输入：半径 ε，给定点在ε 邻域内成为核心对象时邻域内至少要包含的数据对象数 MinPts，数据对象集合 $D = \{x_1, x_2 . \cdots x_n\}$

输出：簇划分C，达到密度要求

REPEAT

从数据库中抽取一个未处理过的点；

IF 抽出的点是核心点 THEN找出所有从该点密度可达的对象，形成一个簇

ELSE 抽出的点是边缘点(非核心对象)，跳出本次循环，寻找下一点；

UNTIL 所有点都被处理；

https://www.bilibili.com/video/BV1KJ411U73w?p=8

https://www.bilibili.com/video/BV1ei4y1x7q1

DBSCAN聚类算法

- Generate clusters of arbitrary shapes.
- Robust against noise.
- No K value required in advance.
- Somewhat similar to human vision.

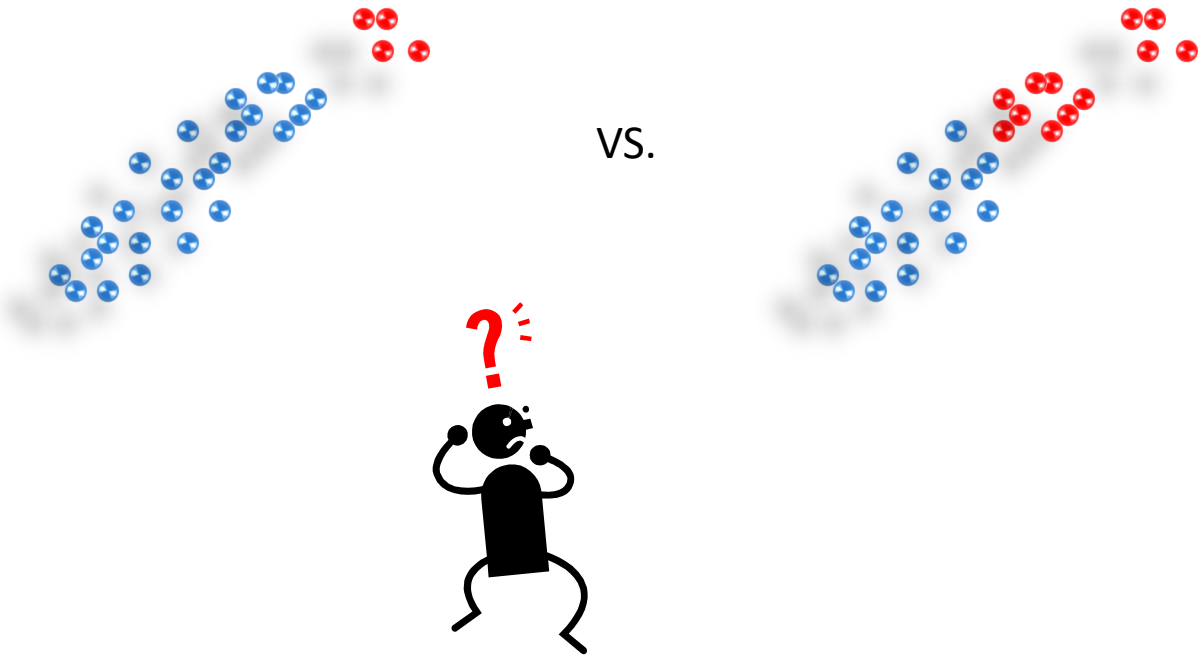
聚类评估

- 内部指标：不需要其他数据

例如 轮廓系数（Silhouette Coefficient），Calinski-Harabaz 指数等

- 外部指标：需要数据真实情况进行对比分析

例如 调整兰德系数（Adjusted Rand index，ARI）等



Silhouette

对于每个样本点 i 的轮廓系数计算为：

$$S(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}}$$

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases}$$

a(i): mean distance between a sample and all other points in the same cluster

b(i): mean distance between a sample and all other points in the next nearest cluster

同cluster样本距离越相近且不同cluster样本距离越远，分数越接近1

实战分析

- DBSCAN密度聚类分析实现
- 聚类分析性能比较

https://www.bilibili.com/video/BV1ST411w7De?p=27

课程总结

- 聚类分析概述
- 聚类方法
  - K均值聚类
  - 层次聚类
  - 密度聚类

Q & A