北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

方以类聚、物以群分

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

开篇语

方以类聚

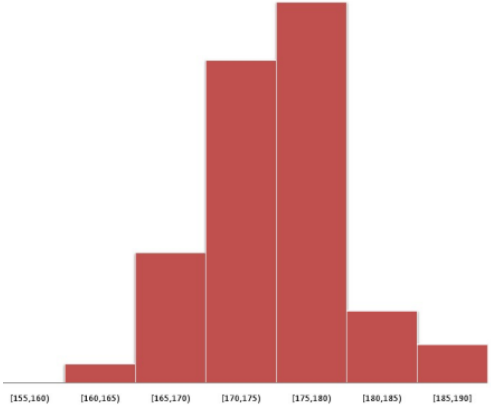
物以群分

数同类者无远

数异类者无近

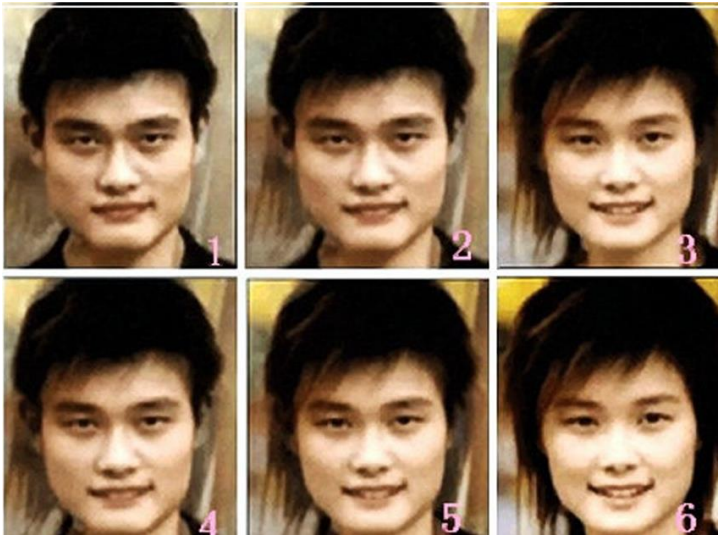
刘徽·《九章算术注·方田九》

什么是聚类



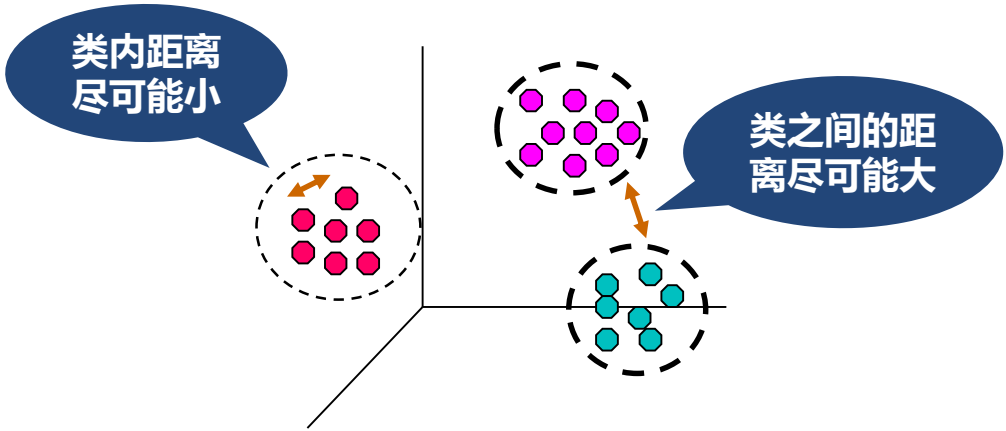
是高是矮，一眼看出两人不是一类人

什么是聚类



头发鼻子脸，多个维度审视之后，才发现他俩如此相似?

什么是聚类

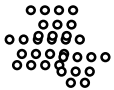


数据有几个变量，就形成几维空间，每个观测值是该空间的一个点。聚类分析就是根据点之间疏密、远近，把它们自然分成不同的簇，聚类结果是簇及其特征

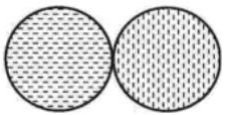
什么是聚类

- 一旦事物离开“类”这一范畴，事物就不能为人所认识和理解
- 族类、物类、类同、类似、不伦不类、……
- 聚类分析cluster analysis简称聚类clustering，是一个把数据对象划分成子集的过程
- 每个子集是一个簇cluster （簇cù），使得簇中的对象彼此相似，但与其它簇中的对象不相似
- 和分类不同，聚类属于无监督学习，即在预先不知道分类的情况下，将数据划分成有意义或有用的簇，捕获数据的自然结构

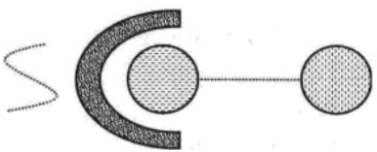
簇的类型



明显分离的簇



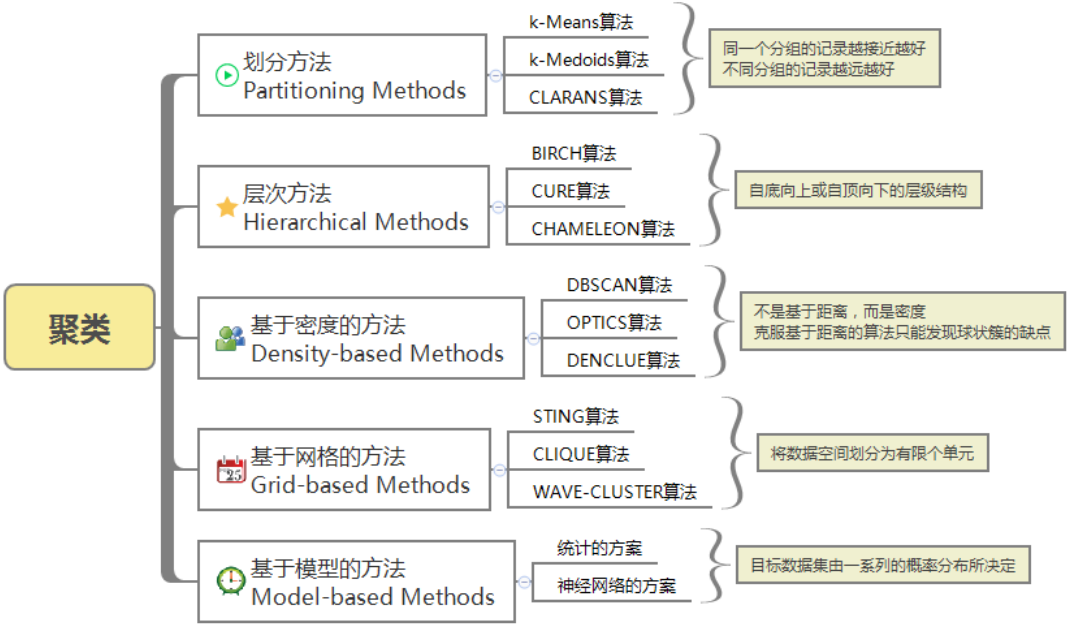
基于中心的簇



基于邻近的簇

基于密度的簇

算法模型


聚类
划分方法
Partitioning Methods
k-Means算法
k-Medoids算法
CLARANS算法
同一个分组的记录越接近越好
不同分组的记录越远越好
层次方法
Hierarchical Methods
BIRCH算法
CURE算法
CHAMELEON算法
自底向上或自顶向下的层级结构
基于密度的方法
Density-based Methods
DBSCAN算法
OPTICS算法
DENCLUE算法
不是基于距离，而是密度
克服基于距离的算法只能发现球状簇的缺点
基于网格的方法
Grid-based Methods
STING算法
CLIQUE算法
WAVE-CLUSTER算法
将数据空间划分为有限个单元
基于模型的方法
Model-based Methods
统计的方案
神经网络的方案
目标数据集由一系列的概率分布所决定

距离计算

若特征均为数值型变量，常采用闵可夫斯基距离：

$$dist\left(x^{(i)},x^{(j)}\right)=\left(\sum_{k=1}^{n}|x_k^{(i)}-x_k^{(j)}|^p\right)^{\frac{1}{p}}$$

对于文本类数据，常采用余弦相似性：$similarity\left(x^{(i)},x^{(j)}\right)=\frac{x^{(i)}\cdot x^{(j)}}{\|x^{(i)}\|\|x^{(j)}\|}$

在R中距离的计算：

若均为数值型变量，用得最多的是 `dist()`实现

对于混合类型数据，可采用`cluster::daisy()`予以实现

数据标准化

为了消除不同单位或量级的影响，在计算距离之前往往要进行标准化：

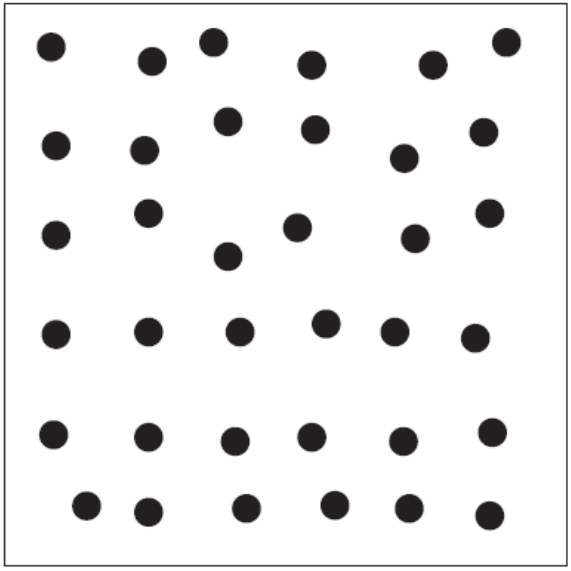
Min-max标准化：$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$

z-score 标准化：$x_i' = \frac{x_i - mean(x_i)}{sd(x_i)}$

正项序列的归一化：$x_i' = \frac{x_i}{sum(x_i)}, x_i > 0$

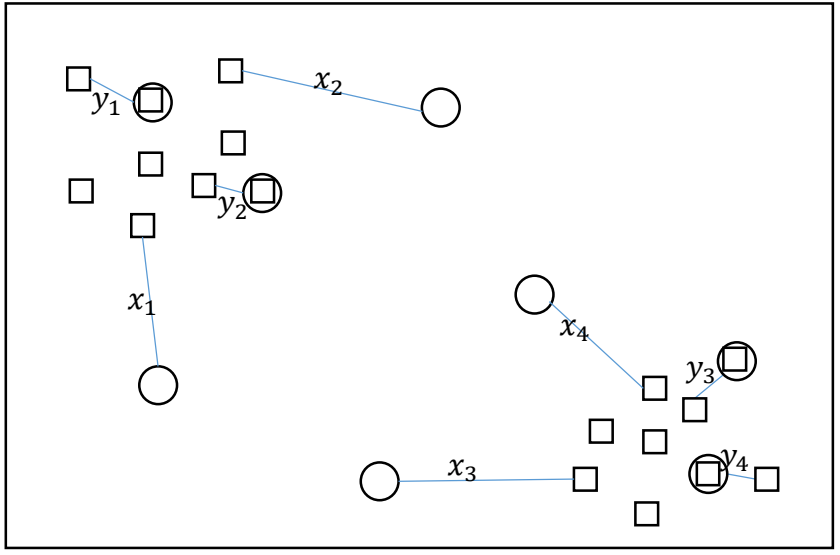
综合而言，不过是：**乘以散之，约以聚之，齐同以通之**

数据是否适合聚类



A data set that is uniformly distributed in the data space

数据是否适合聚类



霍普金斯统计量：

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

若D分布均匀，则H接近于0.5

若D是高度倾斜的，则H接近于0

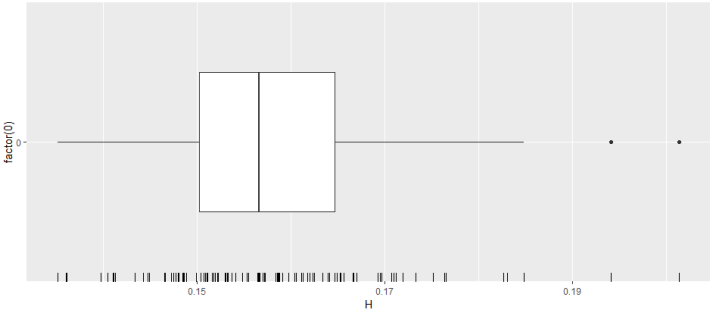
一般而言，$n \ll |D|$

推荐的做法：

$n = 0.05 \times |D|$或是$n = 0.1 \times |D|$

数据是否适合聚类

```
library(clustertend)
set.seed(2012)
scores <- cjb %>%
  select(yw:sw)
n <- floor(nrow(cjb) * 0.05)
hopkins_stat <- unlist(replicate(100,  hopkins(scores, n)))
mean(hopkins_stat)
#> [1] 0.1577968
```



模型评估：轮廓系数

综合考虑凝聚性和分离性，采用轮廓系数silhouette coefficient评估聚类结果：

1：对于第i 个对象，计算它到所属簇中其它所有其它对象的平均距离，记作a_i

2：对于第i 个对象和所有不包含该对象的其它簇，计算该对象到各簇每个对象距离的平均值，并找到不同簇平均值中的最小值，记作b_i

3：对于第i 个对象，轮廓系数为：

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

可以计算所有对象轮廓系数的平均值，得到聚类效果的总体度量

`cluster::silhouette()` compute or extract silhouette information

模型评估：其它方法

`fpc::cluster.stats()` compute several cluster validity statistics from a clustering and a dissimilarity matrix

`clValid::clValid()` calculate validation measures for a given set of clustering algorithms and number of clusters

`cclust::clustIndex()` calculate the values of several clustering indexes, which can be independently used to determine the number of clusters existing in a data set

`NbClust::NbClust()` provide 30 indices for cluster validation and determining the number of clusters

谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics