



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY  
SCHOOL OF ECONOMICS AND MANAGEMENT

# 第 4 讲：聚类分析

康雁飞

数量经济与商务统计系

# Outline

- 1 为什么聚类？
- 2 如何度量相似性？
- 3 如何聚类？
- 4 作业

# Outline

- 1 为什么聚类?
- 2 如何度量相似性?
- 3 如何聚类?
- 4 作业

# 为什么聚类？

- 1 海量数据
- 2 指标多而杂
- 3 数据内在特征不直观

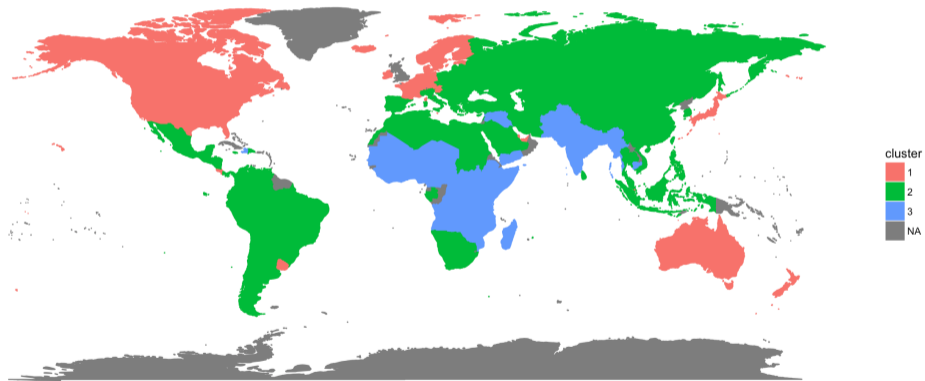


- 1 聚类分析（对数据）
- 2 主成分分析（对指标）

# 举例：幸福指数

## Applied Clustering World Happiness and Social Progress Index

Based on data from:[https://en.wikipedia.org/wiki/World\\_Happiness\\_Report](https://en.wikipedia.org/wiki/World_Happiness_Report) and  
[https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_Social\\_Progress\\_Index](https://en.wikipedia.org/wiki/List_of_countries_by_Social_Progress_Index)




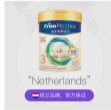










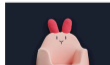


# 举例：顾客行为

淘宝网 Taobao.com

宝贝 > 夏欧美衬衫男

搜索

 <p>卡卡家INS欧美外贸春夏男童帅气加油长袖上衣+迷彩长裤</p> <p>¥35.9 销量:3</p>	 <p>介直 / 儿童小椅子 黑胡桃实木天然木蜡油无异味 家童件 / 芥</p> <p>¥499 销量:28</p>	 <p>简约实木儿童椅子可爱靠背椅家用宝宝小餐椅子板凳幼儿园座椅</p> <p>¥110.6 销量:11</p>	 <p>皇家美素佳儿婴幼儿配方奶粉3段800g3罐</p> <p>HOT ¥951 销量:202</p>	 <p>一家一匠日式儿童学习椅布艺实木可升降可拆洗儿童椅书桌</p> <p>销量:18</p>
 <p>J.LINDBERG金林德伯格夏季新品时尚潮流亚麻短袖衬衫男</p> <p>¥420 销量:2</p>	 <p>卡卡家INS欧美夏季款男宝宝短袖+云朵长裤2件套套装</p> <p>¥33.9 销量:3</p>	 <p>卡卡家INS欧美外贸夏季款女宝宝童短袖豹纹上衣+皮长裤潮款时</p> <p>¥35.9 销量:3</p>	 <p>荷兰美素佳儿进口婴幼儿配方奶粉3段700g2盒</p> <p>限时特卖 1件95折 186 销量:120</p>	 <p>卡卡家INS欧美款秋冬新款男宝宝童龙帅气连帽外套外出服</p> <p>¥43.9 销量:8</p>
				

爱逛好店  
好店直播  
品质特色  
实惠特卖  
猜你喜欢  
顶部  
反馈  
帮助举报

# 举例：文本聚类

## 今日头条

推荐

西瓜视频

热点

直播

图片

科技

娱乐

游戏

体育

懂车帝

财经

搞笑

更多

发微头条 写文章 提问 发视频

有什么新鲜事想告诉大家

0/2000字

图片 表情

发布

### 时政微纪录 | 决战决胜——习近平指挥脱贫攻坚进行时

学习强国

央视网新闻 · 65评论 · 刚刚

### 习近平走秦晋，关注脱贫攻坚三问题

新华网客户端 · 40评论 · 7分钟前



#### 离奇! 时隔28年, 突然想起买过一套房! 上门发现竟住着陌生人

中国青年网 · 427评论 · 15分钟前

### 事情闹大了, 上百名各国政要发联名公开信谴责

环球网 · 229评论 · 22分钟前



#### 贵阳市政协副主席... 九三学社贵阳市委主委田茂书

搜索站内资讯、视频或用户

搜索

登录后可以保存您的浏览喜好、评论、收藏，  
并与APP同步，更可以发布微头条

登录



QQ



微信

## 24小时热闻



#### 工行系王涛接任成都银行行长 原行长王晖留任董事长



#### 复盘2020深圳楼市魔幻开局: 房价暴涨的谎言与真相



#### 金喜爱与刘敏涛谁更杀? 姐姐们重新定义女性美, 女人不为年龄限制

# Outline

- 1 为什么聚类?
- 2 如何度量相似性?
- 3 如何聚类?
- 4 作业



## 相似性度量 - 距离

- 1 在聚类分析中，如果样本点为有限维定量指标，常用明考夫斯基距离 (Minkowski distance)。
- 2 余弦距离（衡量文本之间的相似度通常用）。

# 明考夫斯基距离 (Minkowski distance)

明考夫斯基距离:

$$d_q(x, y) = \left[ \sum_{j=1}^p |x_j - y_j|^q \right]^{1/q}.$$

1  $q = 1$  时, 即绝对值距离:  $d_1(x, y) = \sum_{j=1}^p |x_j - y_j|.$

2  $q = 2$  时, 即欧式距离:  $d_2(x, y) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}.$

3  $q = \infty$  时, 即切比雪夫距离:  $d_\infty(x, y) = \max_{1 \leq j \leq p} |x_j - y_j|.$

# 明考夫斯基距离

- 1 最常用的是欧氏距离。它的优点是：坐标经旋转变换后，点和点之间距离保持不变。
- 2 采用明氏距离时，应采用相同量纲的变量。
- 3 尽可能避免数据的多重相关性。

余弦距离：

$$\cos(x, y) = \frac{x \cdot y}{|x| \cdot |y|}.$$

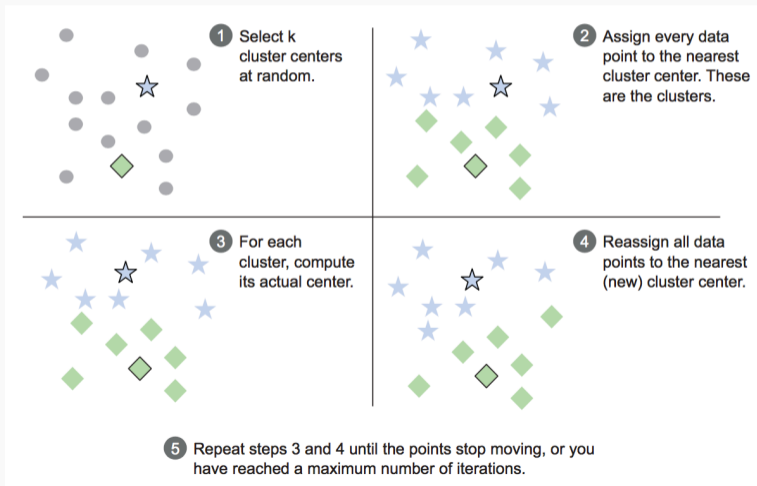
余弦距离可以衡量不同量纲的变量之间的距离（距离：文本）。

# Outline

- 1 为什么聚类?
- 2 如何度量相似性?
- 3 如何聚类?
- 4 作业

# k-均值聚类 (k-means clustering)

k-均值聚类是一个迭代聚类算法。



# k-均值聚类 (k-means clustering)

## k-均值聚类

- 1 确定类的个数  $k$ 。
- 2 随机抽取  $k$  个样本点作为  $k$  个类中心。
- 3 将每个样本点分配到距离其最近的类中。
- 4 重新计算类中心。

重复第 3、4 步直到样本点的类别不再变化或者达到了最大迭代次数。

## 如何选择 $k$ ?

- 如何衡量聚类质量 (goodness of clustering)?
- 组内距越小越好, 组间距越大越好。
- 组内平方和:

$$tot.within = \sum_{k=1}^k \sum_{x_i \in C_k} (x_i - \mu_k)^2,$$

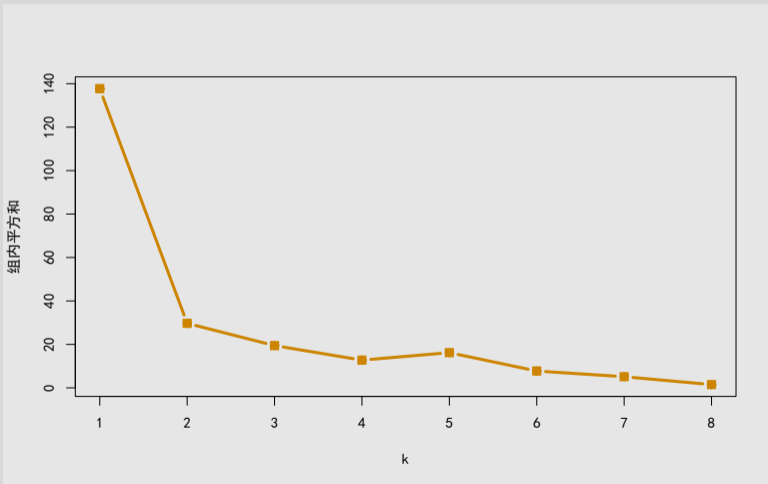
其中  $\mu_k$  为第  $k$  个类  $C_k$  的中心。

- 问题: 组内平方和如何随  $k$  变化?



# 如何选择 $k$ ?

## 碎石图 (Scree plot)



## R 中进行 $k$ -均值聚类

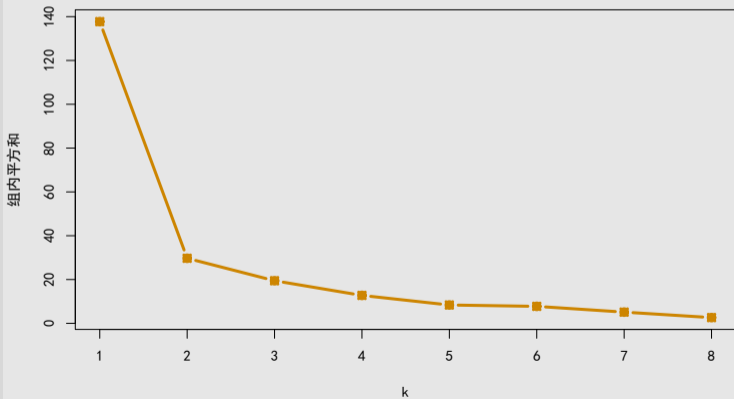
例：某企业生产新式大衣，将新产品的样品寄给九个城市的进货员，并附寄调查意见表征求对新产品的评价，评价分质量、款式、颜色三个方面，以十分评分。

```
coat.scores <- read.csv('./data/coatscores.csv')
str(coat.scores)
```

```
## 'data.frame':  9 obs. of  3 variables:
## $ Quality: int  3 4 10 8 7 3 8 6 9
## $ Style  : num  5 2 7.5 9.5 8 4 6.5 3 8.5
## $ Color  : num  4 5 8.5 7 9 3.5 7 5.5 6
```

# 选择 $k$

## 碎石图 (Scree plot)

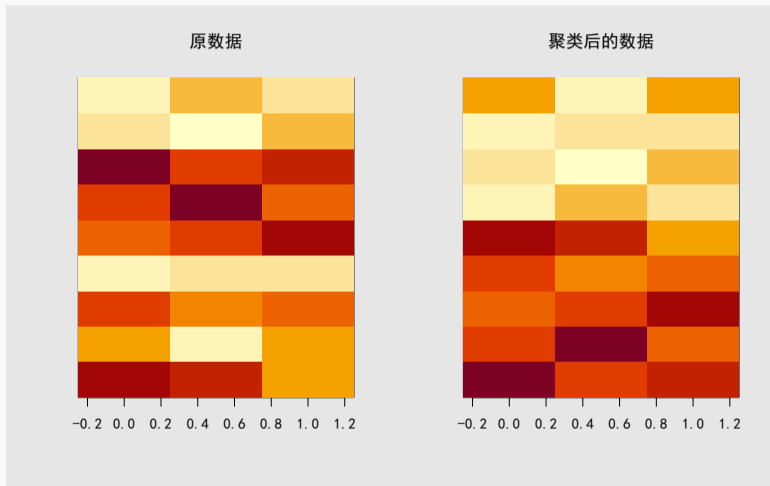


# R 中进行 $k$ -均值聚类

```
coat.scores.kmeans <- kmeans(coat.scores, 2)
coat.scores.kmeans
```

```
## K-means clustering with 2 clusters of sizes 5, 4
##
## Cluster means:
##   Quality Style Color
## 1     8.4   8.0   7.5
## 2     4.0   3.5   4.5
##
## Clustering vector:
## [1] 2 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 16.2 13.5
## (between_SS / total_SS =  78.4 %)
##
## Available components:
```

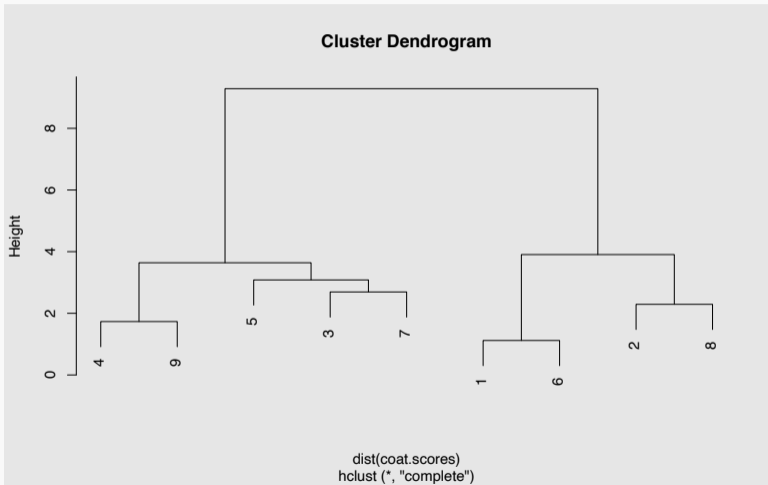
# 可视化



# 分层聚类 (Hierarchical clustering)

- 1 分层聚类算法也是一种迭代算法。
- 2 它不需要我们提前指定分类个数。
- 3 它可以输出一个树状结构。

# 分层聚类



## 分层聚类

- 1 每个样本点自成一类。
- 2 选择最近的两个类聚成一类。
- 3 计算新的类与类之间的距离。

重复第 2、3 步直至所有的样本点聚为一类。



## 如何计算类 $C_1$ 与类 $C_2$ 之间的距离？聚合指数

- 1 最短距离法:  $D(C_1, C_2) = \min_{\substack{x_i \in C_1 \\ y_j \in C_2}} \{d(x_i, y_j)\}.$
- 2 最长距离法:  $D(C_1, C_2) = \max_{\substack{x_i \in C_1 \\ y_j \in C_2}} \{d(x_i, y_j)\}.$
- 3 重心法:  $D(C_1, C_2) = d(\bar{x}, \bar{y}).$
- 4 类平均法:  $D(C_1, C_2) = \frac{1}{l \times m} \sum_{x_i \in C_1} \sum_{y_j \in C_2} d(x_i, y_j).$

## 例：大衣评分

首先需要明确：

- 1 如何测度距离  $d(x_i, y_j)$ ? (欧式距离)
- 2 如何测度聚合指数  $D(C_1, C_2)$ ? (最长距离法)

## 例：大衣评分

```
dist(coat.scores)
```

```
##           1           2           3           4           5           6           7           8
## 2 3.316625
## 3 8.689074 8.860023
## 4 7.365460 8.732125 3.201562
## 5 7.071068 7.810250 3.082207 2.692582
## 6 1.118034 2.692582 9.287088 8.215838 7.889867
## 7 6.020797 6.344289 2.692582 3.000000 2.692582 6.595453
## 8 3.905125 2.291288 6.726812 6.964194 6.184658 3.741657 4.301163
## 9 7.228416 8.261356 2.872281 1.732051 3.640055 7.905694 2.449490 6.284903
```

首先 **1** 和 **6** 合并为一类。

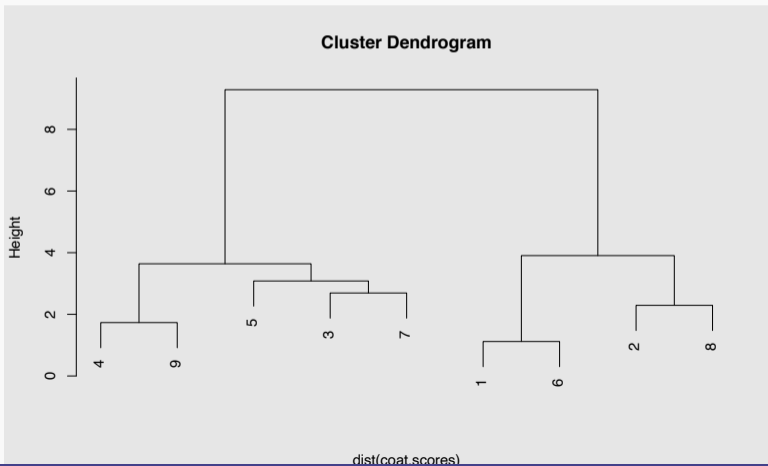
## 例：大衣评分

```
coat.scores.hc$merge
```

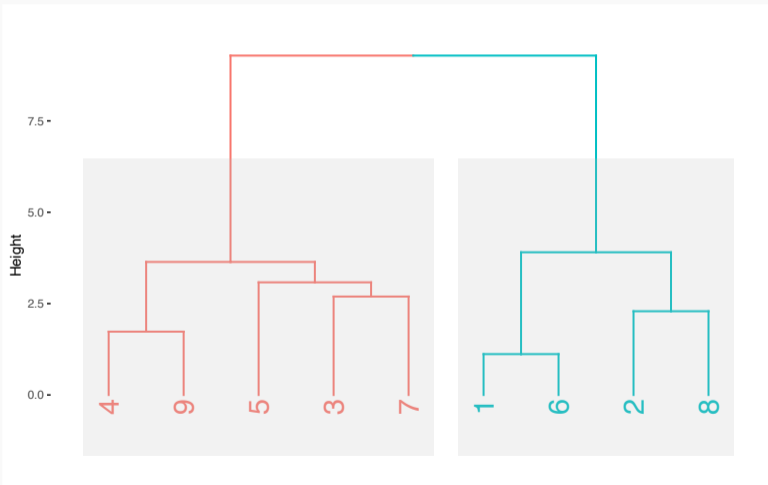
```
##      [,1] [,2]
## [1,]  -1  -6
## [2,]  -4  -9
## [3,]  -2  -8
## [4,]  -3  -7
## [5,]  -5   4
## [6,]   2   5
## [7,]   1   3
## [8,]   6   7
```

# R 中的层次聚类

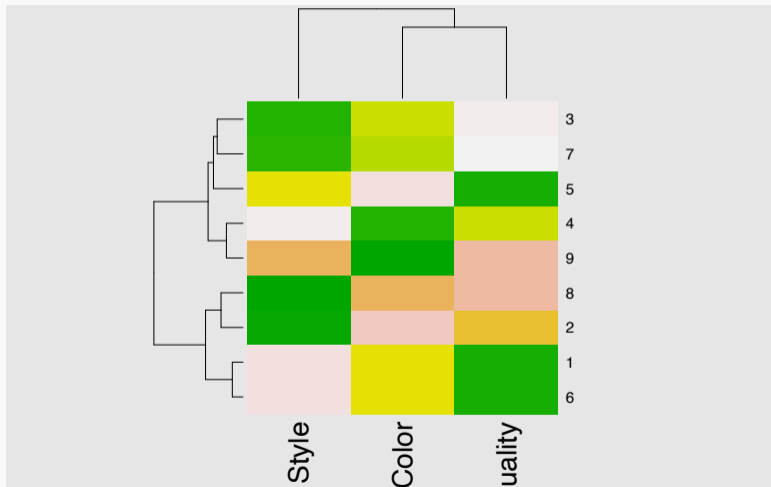
```
coat.scores.hc <- hclust(dist(coat.scores))  
plot(coat.scores.hc)
```



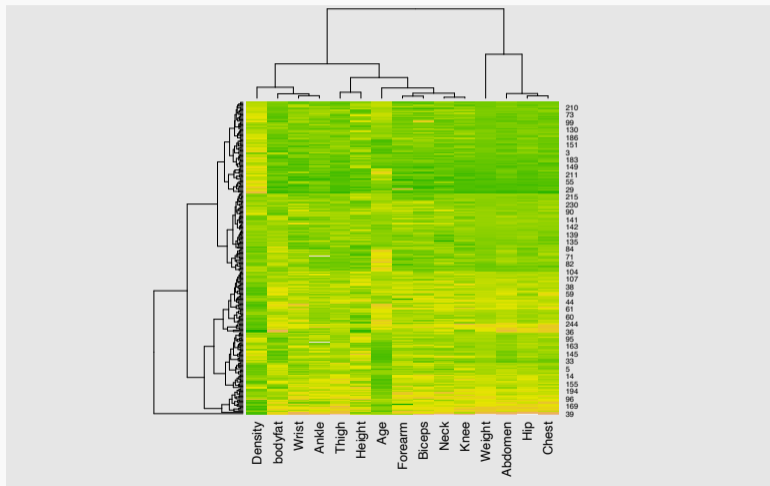
# R 中的层次聚类



# 热力图



# 热力图





- **注意:**至今为止我们理解了样本的聚类。
- 有时候变量聚类非常有用（用户画像）。
- 变量之间的相似性度量：相关系数。
- 变量类之间的聚合指数：最大系数法、最小系数法。

# 其他常用的聚类算法

## 机器学习中的无监督分类方法

- 1 单指标分裂聚类法
- 2 基于密度的聚类算法
- 3 自组织图聚类 (Self-Organized Maps)
- 4 混合模型聚类
- 5 ...

# Outline

- 1 为什么聚类？
- 2 如何度量相似性？
- 3 如何聚类？
- 4 作业

- 1 下表给出六种精神治疗药物的三种临床测量指标数据，请利用分层聚类做聚类分析（分别采用最短距离法和最长距离法）。

药物	吸入量	疗效	依赖性
速可眠	5	9	20
LSD	6	11	2
安定	4	5	20
吗啡	6	9	46
仙人球毒碱	5	7	1
酒精	3	1	12

- 2 请采用聚类分析的方法，对污染数据.xlsx 进行聚类分析。并解释各类的意义。