



北京邮电大学

BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS



Data Analytics with R
语言数据分析



数据对象

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



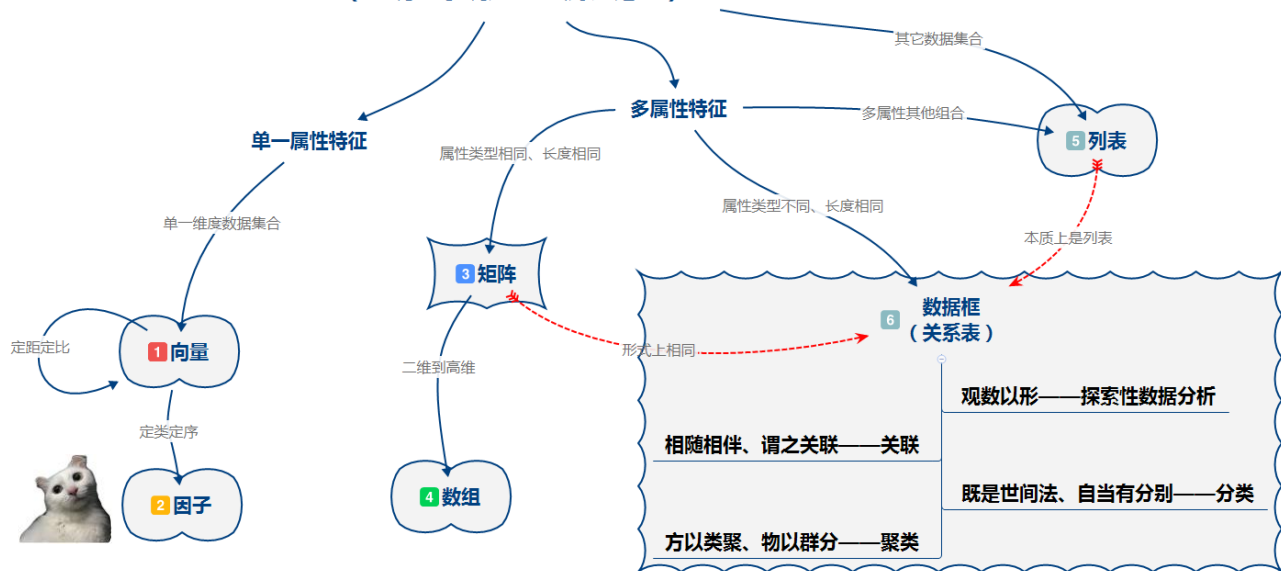
- 第10章 观数以形
- 第11章 相随相伴、谓之关联
- 第12章 既是世间法、自当有分别
- 第13章 方以类聚、物以群分
- 第14章 庐山烟雨浙江潮

数据对象



当前位置

设备或人工采集到的数据
(对现实系统进行观测、记录)




测量的尺度

类型	量化程度	举例	数学性质	R数据对象
定类 nominal	是否相同	性别、人种	$=$ \neq	无序因子
定序 ordinal	比较大小	等级、规模	$>$ $<$	有序因子
定距 interval	确定差别	温度、时刻	$+$ $-$	数值向量
定比 ratio	确定比例	长度、薪资	\times \div	数值向量


更多内容请参阅: Stevens, Stanley Smith. *On the theory of scales of measurement.*
Science 1946: 677-680.

因子 (Factor)

 向量用于存储数值变量（定距定比），因子用于存储类别变量（定类定序）

 作为类别变量，只有有限个取值（类别），称为水平levels，取值水平往往远远少于观测对象（记录）的个数

 因子也是更有效的存储方式：存储为整型向量，只不过每一个1~nlevels的正整数代表了相应的类别

 在分组统计中，因子常用来作分组变量；分类问题均要求因变量为因子；在其它一些算法建模过程中，也要求其变量为因子（如apriori算法）

因子的创建

```
xb <- c("女", "男", "男", "女", "男", "女")
```

```
xb
```

```
#> [1] 女 男 男 女 男 女
```

```
typeof(xb)
```

```
#[1] "character"
```

```
xb <- factor(xb)
```

```
xb
```

```
#> [1] 女 男 男 女 男 女
```

```
#> Levels: 男 女
```

因子的基本操作

```
xb[c(1, 4:5)]
```

```
#> [1] 女 女 男
```

```
#> Levels: 男 女
```

```
xb[-c(2:3, 6)]
```

```
#> [1] 女 女 男
```

```
#> Levels: 男 女
```

```
xb[1] <- "男"
```

```
xb
```

```
#> [1] 男 男 男 女 男 女
```

```
#> Levels: 男 女
```

```
xb == "男"
```

```
#> [1] TRUE TRUE TRUE FALSE TRUE FALSE
```

因子的基本操作

`nlevels(xb)` #取值水平的个数

```
#> [1] 2
```

`levels(xb)` #取值水平, 其顺序可参阅?Comparison的结果

```
#> [1] "男" "女"
```

```
xb[1] <- "中性"
```

```
#> Warning message:
```

```
#> In `[<-.factor`(`*tmp*`, 1, value = "中性") :
```

```
#> invalid factor level, NA generated
```


因子的基本操作

```
xb <- c("女", "男", "男", "女", "男", "女")
xb <- factor(xb, levels = c("男", "女", "中性"))
xb
#> [1] 女 男 男 女 男 女
#> Levels: 男 女 中性
table(xb)
#> xb
#> 男    女  中性
#> 3     3     0
xb[1] <- "中性" #此时可以赋值了
xb
#> [1] 中性 男    男    女    男    女
#> Levels: 男 女 中性
```

因子的基本操作

```
typeof(xb)
```

```
#[1] "integer"
```

```
as.numeric(xb)
```

```
#> [1] 2 1 1 2 1 2
```

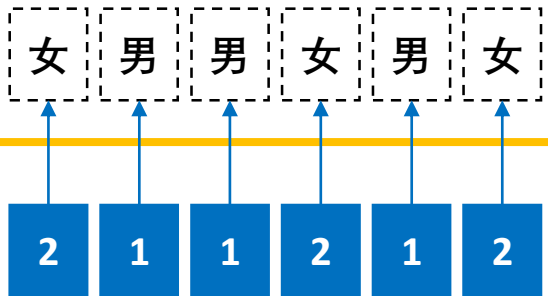
```
as.character(xb)
```

```
#> [1] "女" "男" "男" "女" "男" "女"
```

面子



里子



因子的基本操作

```
number_factors <- factor(c(10,20,20,20,10))
mean(number_factors)
#> [1] NA
mean(as.numeric(number_factors))
#> [1] 1.6
as.numeric(number_factors)
#> [1] 1 2 2 2 1
mean(as.numeric(as.character(number_factors)))
#> [1] 16
mean(as.numeric(levels(number_factors)[number_factors]))
#> [1] 16
```

有序因子

#男女平等，xb为无序因子，因而下述逻辑运算符没有意义

```
xb[1] > xb[2]
```

```
#> [1] NA
```

```
#> Warning message:
```

```
#> In Ops.factor(xb[1], xb[2]) : '>' not meaningful for factors
```

```
score <- factor(c("优", "良", "优", "优", "良", "优"),  
                ordered = TRUE)
```

```
score[1] > score[2]
```

```
#> [1] TRUE
```

有序因子

```
days <- factor(c("周一", "周三", "周二", "周二"),  
               ordered = TRUE)
```

```
days[3] < days[2]
```

```
#> [1] TRUE
```

```
days[1] < days[2]
```

```
#> [1] FALSE
```

```
days
```

```
#> [1] 周一 周三 周二 周二
```

```
#> Levels: 周二 < 周三 < 周一
```

有序因子

```
days <- factor(c("周一", "周三", "周二", "周二"),  
               ordered = TRUE,  
               levels = c("周一", "周二", "周三"))
```

```
days
```

```
#> [1] 周一 周三 周二 周二
```

```
#> Levels: 周一 < 周二 < 周三
```

```
days[3] < days[2]
```

```
#> [1] TRUE
```

```
days[1] < days[3]
```

```
#> [1] TRUE
```

实战才是王道：数据分箱

#百分制成绩变为五分制成绩

```
yw <- c(94, 87, 92, 91, 85, 92)
```

#数据分箱

```
yw5 <- cut(yw,  
            breaks = c(0, (6:10)*10))
```

yw5

```
#> [1] (90,100] (80,90] (90,100] (90,100] (80,90] (90,100]
```

```
#> Levels: (0,60] (60,70] (70,80] (80,90] (90,100]
```

实战才是王道：数据分箱

#百分制成绩变为五分制成绩

```
yw <- c(94, 87, 92, 91, 85, 92)
```

#数据分箱+闭区间

```
yw5 <- cut(yw,  
            breaks = c(0, (6:10)*10) ,  
            include.lowest = TRUE)
```

yw5

```
#> [1] (90,100] (80,90] (90,100] (90,100] (80,90] (90,100]
```

```
#> Levels: [0,60] (60,70] (70,80] (80,90] (90,100]
```


实战才是王道：数据分箱

#百分制成绩变为五分制成绩

```
yw <- c(94, 87, 92, 91, 85, 92)
```

#数据分箱+闭区间+左开右闭

```
yw5 <- cut(yw,  
            breaks = c(0, (6:10)*10) ,  
            include.lowest = TRUE,  
            right = FALSE)
```

yw5

```
#> [1] [90,100] [80,90) [90,100] [90,100] [80,90) [90,100]
```

```
#> Levels: [0,60) [60,70) [70,80) [80,90) [90,100]
```

实战才是王道：数据分箱

#百分制成绩变为五分制成绩

```
yw <- c(94, 87, 92, 91, 85, 92)
```

#数据分箱+闭区间+左开右闭+有序因子

```
yw5 <- cut(yw,  
            breaks = c(0, (6:10)*10) ,  
            include.lowest = TRUE,  
            right = FALSE,  
            ordered_result = TRUE)
```

yw5

```
#> [1] [90,100] [80,90) [90,100] [90,100] [80,90) [90,100]
```

```
#> Levels: [0,60) < [60,70) < [70,80) < [80,90) < [90,100]
```

实战才是王道：数据分箱

#百分制成绩变为五分制成绩

```
yw <- c(94, 87, 92, 91, 85, 92)
```

#数据分箱+闭区间+左开右闭+有序因子+标签

```
yw5 <- cut(yw,  
            breaks = c(0, (6:10)*10) ,  
            include.lowest = TRUE,  
            right = FALSE,  
            ordered_result = TRUE,  
            labels = c("不及格", "及格", "中", "良", "优"))
```

```
yw5
```

```
#> [1] 优 良 优 优 良 优
```

```
#> Levels: 不及格 < 及格 < 中 < 良 < 优
```

A decorative blue border with rounded corners frames the entire slide. Two thin blue crosshair symbols are positioned on the right and left sides of the text.

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>

