数据科学导论第 4 讲-数据分析基础

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 03 月 31 日

中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

目录

数据处理

数据清洗

数据变换

实际操作

数据处理

数据分类

- 数据是数据对象及其属性的集合。一个数据对象是对一个事物或者物理对象的描述，一个典型的数据对象可以是一条记录、一个实体、一个案例或一个样本等等。
- 数据对象的属性则是这个对象的性质或特征，例如一个人的肤色、眼球颜色是这个人的属性，而某地某天的气温则是该地该天气象记录的属性特征。
- 大数据时代，数据的来源越来越多样化，比如来自互联网、银行、工商、税务、公安天眼等等。同时，数据的格式和形态也越来越多样化，有数字、文字、图片、音频、视频等等。

数据分类

- 能够用统一的结构加以表示的数据，如数字、符号等，我们称之为结构化数据；无法用统一的结构表示的数据，如**文本**、**音频**、**图像**、**视频**，我们称之为非结构化数据。
- 过去所分析的数据大部分是**结构化数据**，但是随着非结构化数据越来越多，有必要去研究非结构化数据。

数据类型和特征

- 对于结构化数据，按照对客观事物测度的程度或精确水平来划分，可将数据的计量尺度从低级到高级、由粗略到精确划分为四种，如表 1 所示。

数据类型

数据类型	数据特征	举例
分类数据（categorical data）	没有数量关系，没有顺序关系	状态，如‘男’‘女’、‘0’‘1’
有序数据（ordinal data）	有顺序关系	特征量，如‘甲’‘乙’‘丙’‘丁’、甲>乙>丙>丁
区间数据（interval data）	有数量关系，可比较大小，可排序，可计算差异	实数，如长度、重量、压力
比例数据（ratio data）	实数，事物之间的比值 有数量关系，可以比较大小，可排序，可计算差异，具有绝对零点	实数，事物之间的比值

图 1: 常见的数据类型及其特征

数据分类

- 在计量尺度的应用中，需要注意的是，同类事物用不同的尺度量化，会得到不同的类别数据。
- 如农民收入数据按实际填写就是区间数据；按高、中、低收入水平分就是有序；按有无收入计量则是分类；而说某人的收入是另一人的两倍，便是比例数据了。

数据清洗

数据清洗 (Data Cleaning)

1. 数据清洗是数据准备过程中最重要的一步。
2. 通过填补**缺失数值、光滑噪声数据、识别或删除离群点**并解决不一致性来“清洗”数据，进而达到数据格式标准化，清除**异常数据、重复数据，纠正错误数据**等目的。

缺失数据处理

- 从缺失的分布来讲，缺失值可以分为完全随机缺失（missing completely at random, MCAR），随机缺失（missing at random, MAR）和完全非随机缺失（missing not at random, MNAR）。

```
a  <- c(1,2,3,4,NA,5,3,2,3,4,NA)
is.na(a)
```

```
##  [1] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FA
```

```
sum(a,na.rm=TRUE)
```

```
## [1] 27
```

缺失数据

```
na.omit(a)
```

```
## [1] 1 2 3 4 5 3 2 3 4
## attr(,"na.action")
## [1]  5 11
## attr(,"class")
## [1] "omit"
```

```
b <- na.omit(a)
print(b)
```

```
## [1] 1 2 3 4 5 3 2 3 4
## attr(,"na.action")
## [1]  5 11
## attr(,"class")
## [1] "omit"
```

缺失数据

1. 完全随机缺失是指数据的缺失是完全随机的，不依赖于任何完全变量或不完全变量。缺失情况相对于所有可观测和不可观测的数据来说，在统计意义上是独立的，也就是说直接删除缺失数据对建模影响不大。
2. 随机缺失指的是数据的缺失不是完全随机的，数据的缺失依赖于其他完全变量。具体来说，一个观测出现缺失值的概率是由数据集中不含缺失值的变量决定的，与含缺失值的变量关系不大。
3. 完全非随机缺失指的是数据的缺失依赖于不完全变量，与缺失数据本身存在某种关联，比如调查时，所设计的问题过于敏感，被调查者拒绝回答而造成的缺失。

缺失数据

从**统计角度**来看，非随机缺失的数据会产生有偏估计，而非随机缺失数据处理也是比较困难的。

事实上，绝大部分的原始数据都包含有缺失数据，因此怎样处理这些缺失值就很重要了。

缺失数据

- 在 R 中，缺失值以符号 **NA** 表示。Python, NaN
- 可以使用赋值语句将某些值重新编码为缺失值，例如：

```
a[which(a == 4)] <- NA
print(a)#a
```

```
##  [1]  1  2  3 NA NA  5  3  2  3 NA NA
```

- 任何等于 4 的值都将被修改为 NA。
- 在进行数据分析前，要确保所有的缺失数据被编码为缺失值，否则分析结果将失去意义。

缺失数据

- complete.cases() 函数可用来识别矩阵或数据框的行是否完整的，也就是有无缺失值，返回结果是逻辑值，以行为单位返回识别结果。
- 如果一行中不存在缺失值，则返回 TRUE；若行中有一个或多个缺失值，则返回 FALSE。由于逻辑值 TRUE 和 FALSE 分别等价于数值 1 和 0，可用 sum() 和 mean() 来计算关于完整数据的行数和完整率。

```
complete.cases(a)
```

```
##  [1]  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FA
```

缺失数据

```
sum(a)
```

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

```
sum(a,na.rm = TRUE)
```

```
## [1] 19
```

```
mean(a,na.rm = TRUE)
```

```
## [1] 2.714286
```

例子

```
data(sleep,package = "VIM" ) # 读取 VIM 包中的 sleep 数据
sleep [!complete.cases (sleep),][1:3,2:4] # 提取 sleep 数据中不完
```

```
##    BrainWgt NonD Dream
## 1    5712.0   NA    NA
## 3      44.5   NA    NA
## 4       5.7   NA    NA
```

```
sum(!complete.cases(sleep))
```

```
## [1] 20
```

```
mean (!complete.cases(sleep))
```

```
## [1] 0.3225806
```

均值插补法（Mean Imputation）

- 如果缺失数据是数值型的，根据该变量的平均值来填充缺失值；如果缺失值是非数值型的，就根据该变量的众数填充缺失值。
- 均值插补法是一种**简便**、**快速**的缺失数据处理方法。使用均值替换法插补缺失数据，对该变量的均值估计不会产生影响。
- 该方法是建立在完全随机缺失的假设之上的，当缺失比例较高时会低估该变量的方差，会产生有偏估计。

均值插补

```
a
```

```
##  [1]  1  2  3 NA NA  5  3  2  3 NA NA
```

```
a[is.na(a)] <- mean(a,na.rm  = TRUE)
a
```

```
##  [1] 1.000000 2.000000 3.000000 2.714286 2.714286 5.000000 3.
##  [9] 3.000000 2.714286 2.714286
```

多重插补（Multiple Imputation，MI）

- 在面对复杂的缺失值问题时，MI 是最常用的方法，它将从一个包含缺失值的数据集中生成一组完整的数据集。
- 每个模拟的数据集中，缺失数据将用**蒙特卡洛方法**来填补。
- 由于多重插补方法并不是用单一值来替换缺失值，而是试图产生缺失值的一个随机样本，反映出了由于数据缺失而导致的不确定。
- R 中的 mice 包 [Multivariate Imputation by Chained Equations] 可以用来多重插补。

噪声数据

- 数据噪声是指数据中存在的随机性错误或偏差，产生的原因很多。
- 噪声数据的处理方法通常有分箱、聚类分析和回归分析等，有时也会将与人的经验判断相结合。
- 分箱是一种将数据排序并分组的方法，分为等宽分箱和等频分箱。
- 等宽分箱，是用同等大小的格子来将数据范围分成 N 个间隔。
- 等宽分箱比较直观和容易操作，但是对于偏态分布的数据，等宽分箱并不是太好，因为可能出现许多箱中没有样本点的情况。
- 等频分箱是将数据分成 N 个间隔，每个间隔包含大致相同的数据样本个数，这种分箱方法有着比较好的可扩展性。将数据分箱后，可以用箱均值、箱中位数和箱边界来对数据进行平滑，平滑可以在一定程度上削弱离群点对数据的影响。

噪声数据

- R 语言的等宽分箱法一般都是用 cut 来获取，把连续数列，根据等宽分箱的办法切分开来。

```
d <- c(1,2,3,4,5,6,4,3,2,1)
cut(d,10)
```

```
##  [1] (0.995,1.5] (1.5,2]     (2.5,3]     (3.5,4]     (4.5,5]
##  [7] (3.5,4]     (2.5,3]     (1.5,2]     (0.995,1.5]
## 10 Levels: (0.995,1.5] (1.5,2] (2,2.5] (2.5,3] (3,3.5] (3.5,4
```

```
cut(d,10,labels=F)# 打标签
```

```
##  [1]  1  2  4  6  8 10  6  4  2  1
```

```
d[cut(d,10,labels=F)==10]# 获取标签 10 的数据
```

```
## [1] 6
```

噪声数据

- 聚类分析处理噪声数据是指先对数据进行聚类，然后使用聚类结果对数据进行处理，如舍弃离群点、对数据进行平滑等。类似于分箱，可以采用中心点平滑、均值平滑等方法来处理。
- 回归分析处理噪声数据是指对于利用数据建立回归分析模型，如果模型符合数据的实际情况，并且参数估计是有效的，就可以使用回归分析的预测值来代替数据的样本值，降低数据中的噪声和离群点的影响。

异常值处理

- 常用的异常值处理操作包括 BOX-COX 转换（处理有偏分布），箱线图分析删除异常值，长尾截断等方式，当然这些操作一般都是处理数值型的数据。
- 一般是用于连续的变量不满足正态的时候，在做线性回归的过程中，一般线性模型假定：

$$Y = X\beta + \varepsilon$$

- 其中 ε 满足正态分布，但是利用实际数据建立回归模型时，个别变量的系数通不过。
- 例如往往不可观测的误差 ε 可能是和预测变量相关的，不服从正态分布，于是给线性回归的最小二乘估计系数的结果带来误差，为了使模型满足**线性性、独立性、方差齐性以及正态**性，需改变数据形式，故应用 BOX-COX 转换。

转换非正态数据分布的方式

1. 对数转换：$y_i = ln(x_i)$
2. 平方根转换：$y_i = \sqrt{(x_i)}$
3. 倒数转换：$y_i = 1/x_i$
4. 平方根后取倒数：$y_i = 1/\sqrt{x_i}$
5. 平方根后再取反正弦：$y_i = arcsin(\sqrt{x_i})$
6. 幂转换：$y_i = (x_i^{\lambda} - 1)/(\widetilde{x}^{\lambda+1})$，其中 $\widetilde{x} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$，且参数 $\lambda \in [-1.5, 1]$

在一些情况下（P 值 <0.003）上述方法很难实现正态化处理，所以优先使用 BOX-COX 转换，但是当 P 值 >0.003 时两种方法均可，优先考虑普通的平方变换。

BOX-COX 的变换公式

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+c)^{\lambda}}{\lambda}, \lambda \neq 0 \\ log(y+c), \lambda = 0 \end{cases}$$

数据变换

数据变换 (Data Transformation)

- 数据变换包括平滑、聚合、泛化、规范化、属性和特征的重构等操作。
  1. 数据平滑：指的是将噪声从数据中移出。
  2. 数据聚合：数据聚合指的是将数据进行汇总，以便于对数据进行统计分析。
  3. 数据泛化：数据泛化是将数据在概念层次上转化为较高层次的概念的过程。例如，将分类替换为其父分类。数据泛化的主要目的是减少数据的复杂度。
  4. 数据规范化

常用方法

- 标准差标准化，将变量的各个记录值减去其平均值，再除以其标准差，即

$$x_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i}$$

其中

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, S_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}$$

。

- 也称 z-score 标准化，经过标准差标准化处理后的数据的平均值为 0，标准差为 1。
- z-score 标准化方法适用于属性 A 的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况。

常用方法

- min-max 标准化是将各个记录值减去记录值的最小值，再除以记录值的极差，即

$$x_{ij} = \frac{x_{ij} - min(x_{ij})}{max(x_{ij}) - min(x_{ij})}$$

- 也叫离差标准化，是对原始数据的线性变换，使结果映射到 [-1,1] 区间。

常用方法

- 比例法（归一化方法），对**正向序列** $x_1, x_2, \cdots, x_n$ 进行变换：

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

- 新序列 y_i 取值范围 [0,1]，且 $\sum_{i=1}^{n} y_i = 1$

实际操作

R 语言

- 参考 R 语言实战（第 2 版），前三章（数据类型和基本操作）

作业

- R 复现：第 4 章
- 要求：将该 4 章正文中代码复现
- 提交：2025 年 4 月 5 日 24 点，畅课平台，R+ 学号 + 姓名（压缩）

本周推荐

1. 一本书：《赫伯特西蒙自传 Models of mylife》，中译出版社
2. 一部电影：《Hidden Figures(隐藏人物)》，2016