第七章  数据挖掘与R语言

1. 简述R语言数据类型和数据结构

**R语言中的数据类型**

**1. 数值型（numeric）：**数据的内容为数字。上图中，定量变量和定性变量都可以用数值表示。下面的例子中，x, y, z, w 均为数值型数据。

x <- 175.3 #设 x为身高，x为定量变量（连续型）y <- 5 #设 y为家庭人口数，y为定量变量（离散型）z <- 6 #设 z为教育程度，6表示本科及以上，z为定性变量（有序）w <- 1 #设 w为性别，1表示女性，w为定性变量（无序）

**2. 字符型（character）：**数据的内容为字符。字符型数据可用来表示定性变量，但不能表示定量变量。只要将内容放入英文双引号 "" 中，该数据即会被R识别为字符型。下面例子中的z, w 均为字符型数据。

z <- "本科及以上" #设 z为教育程度，z为定性变量（有序）w <- "女" #设 w为性别，w为定性变量（无序）

**3. 逻辑型（logical）：**仅有两个取值，TRUE和FALSE，注意必须是大写。

**4. 因子型（factor）：**因子是针对定性变量而言的，刚刚讲到定性变量既可以用数值、也可以用字符表示，在此基础上做一个简单的处理就会成为因子型数据。这个处理不会对数据的内容造成任何改变，但会有助于后续的统计分析工作，之后会详细讲。

**R语言中的数据结构**在刚开始接触统计的时候，我们会经常强调一对概念——总体和样本。但是，这个问题在做回归时可能会被忽略。

初学者们通常会被向量、数组之类的名词搞得一头雾水，其实这些都是表达数据结构的名词，本质就是数据的组合形式。下图展示了R中5种数据结构。

**(a) 向量（vector）**就是一连串数据的组合，可以看做是一行或一列数据，其中的数据类型可以是数值型、字符型、逻辑型或因子型。注意，单个向量中的数据必须拥有相同的类型。

比如上图(a)中的三个小方块可以是1, 20, 100这三个数字，也可以是"小学", "初中", "大学"这三个字符，或是TRUE, FALSE, FALSE这样的逻辑型数据。

**(b) 矩阵（matrix）**是具有一定行数和列数的数据集合。其数据类型可以是数值型、字符型、逻辑型或因子型。矩阵中所有数据的类型必须相同。

**(c) 数组（array）**是矩阵的推广，即在矩阵拥有的两个维度（行、列）的基础上增加了第三个维度。其中的数据也只能拥有一种类型。该类数据结构在一般的统计分析中不常用。

**(d) 数据框（data frame）**的结构类似于矩阵，但它可包含多种数据类型（数值型、字符型、逻辑型或因子型），是最常用的数据结构。通常，数据框中的行表示观察对象（也叫观测/observation），列表示变量（variable）。

**(e) 列表（list**）像一个大抽屉，可以将若干（可能无关的）数据信息整合到单个数据结构中。这里的数据信息可以是包括列表在内的五种数据结构中的任意一种或几种。在R中，由于许多函数的运行结果都是以列表的形式返回的，因此该类数据结构也是学习的重点。

2.简述R语言数据管理常用方法

**可以有多种选择方法**

#选择year列

select(flights, year)

#选择1到3列

select(flights, 1:3)

#选择year到day列

select(flights, year:day)

#选择包含 time 字符的列

select(flights, contains("time"))

#选择s开头的列

select(flights,starts\_with("s"))

#反选，在选择前面加 -

select(flights, -(1:6))

select(flights, -(year:day))

第八章  数据挖掘模型与应用

1. 简述数据挖掘相关概念及常用方法和模型

数据挖掘方法可以粗分为:统计方法、机器学习方法、神经网络方法和数据库方法。统计方法可细分为:回归分析、判别分析等。机器学习可细分为:遗传算法等。神经网络方法可细分为:前向神经网络、自组织神经网络等。数据库方法主要是多维数据分析方法等。



2.简述数据预处理：数据青洗，数据集成，数据规范，数据重构的方法

**数据预处理的方法有：数据清理、 数据集成 、数据规约和数据变换。**

1、数据清洗

数据清洗是通过填补缺失值，平滑或删除离群点，纠正数据的不一致来达到清洗的目的。简单来说，就是把数据里面哪些缺胳膊腿的数据、有问题的数据给处理掉。总的来讲，数据清洗是一项繁重的任务，需要根据数据的准确性、完整性、一致性、时效性、可信性和解释性来考察数据，从而得到标准的、干净的、连续的数据。

（1）缺失值处理

实际获取信息和数据的过程中，会存在各类的原因导致数据丢失和空缺。针对这些缺失值，会基于变量的分布特性和变量的重要性采用不同的方法。若变量的缺失率较高（大于80%），覆盖率较低，且重要性较低，可以直接将变量删除，这种方法被称为删除变量。

若缺失率较低（小于95%）且重要性较低，则根据数据分布的情况用基本统计量填充（最大值、最小值、均值、中位数、众数）进行填充，这种方法被称为缺失值填充。对于缺失的数据，一般根据缺失率来决定“删”还是“补”。

（2）离群点处理

离群点（异常值）是数据分布的常态，处于特定分布区域或范围之外的数据通常被定义为异常或噪声。我们常用的方法是删除离群点。

（3）不一致数据处理

实际数据生产过程中，由于一些人为因素或者其他原因，记录的数据可能存在不一致的情况，需要对这些不一致数据在分析前进行清理。例如，数据输入时的错误可通过和原始记录对比进行更正，知识工程工具也可以用来检测违反规则的数据。

2、数据集成

随着大数据的出现，我们的数据源越来越多，数据分析任务多半涉及将多个数据源数据进行合并。数据集成是指将多个数据源中的数据结合、进行一致存放的数据存储，这些源可能包括多个数据库或数据文件。在数据集成的过程中，会遇到一些问题，比如表述不一致，数据冗余等，针对不同的问题，下面简单介绍一下该如何处理。

（1）实体识别问题

在匹配来自多个不同信息源的现实世界实体时，如果两个不同数据库中的不同字段名指向同一实体，数据分析者或计算机需要把两个字段名改为一致，避免模式集成时产生的错误。

（2）冗余问题

冗余是在数据集成中常见的一个问题，如果一个属性能由另一个或另一组属性“导出”，则此属性可能是冗余的。

（3）数据值的冲突和处理

不同数据源，在统一合并时，需要保持规范化，如果遇到有重复的，要去重。