Вероятности и статистика с R

Асен Чорбаджиев

November 2, 2017

1 Функции за работа със структурирани данни

За работа със структурирани данни в R с цел ефективна обработка на структури от данни се налага избягването на използването на цикли. Вместо тях съществува специален клас функции за тази цел. Те се използват за агрегиране, пресмятания или специфично дефинирани операции върху подмножества или всички налични данни.

1.1 Селектиране и групиране на данни

Отделянето на подмножество в R е възможно чрез:

- Избор на елементи и променливи (Keeping). Достъп с индекси или *%in%*.
- Изключване на елементи или променливи (Dropping). Достъп с отрицателен индекс или !:
  vars <- names(data) %in% c("v1", "v2", "v3")
  newdata <- data[!vars]
- Селектиране: which() или използването на функцията subset( )
- Добавяне на вектор след определен индекс във вектор с функцията append().

1.2 Joint

За комбиниране на data.frame се използва функцията merge(), която идентифицира и комбинира таблиците по съвпадения на колони и редове. Нейната функционалност

наподобява Join операциите в SQL. Селектирането по коя колона се сортира става чрез подаване на ключ чрез by=c(x1,x2,...). Когато общи id имат различни имена се използват by.x и by.y за указване на тяхното съвпадение. R запазва името на първата таблица (by.x). Следните операции са възможни:

- *all=TRUE* : outer join - запазва всички колони и групира левите, които имат съвпадение в дясно
- *all=FALSE* : inner join- само съвпадащите колони от data frames.
- *all.x=TRUE*: left outer join - включва всички колони от лявата таблица и тези от дясно, които съвпадат
- *all.y=TRUE*: right outer join - включва всички колони от дясната таблица и тези от ляво, които съвпадат

2 Операции върху таблици

За ефективен достъп и обрабока на струтурирани данни в R се използва специален клас функции. Те позволяват достъп и изпълнение на предварително дефинирана функция до елементите на вектор, list или data.frame. Такива функции са:

- **apply(array, margin, function, ...)** : Връща вектор от резултата на function върху margin.
  mat <- matrix(rep(seq(4), 4), ncol = 4)
  apply(mat, 1, sum) # редове
  apply(mat, 2, sum)
  apply(mat1, 1, function(x) sum(x) + 2)
- **lapply(list, function, ...)** Прилага function върху вектор или лист. Връща list. Особено полезна за работа с data.frame.
  mat.df <- data.frame(mat)
  lapply(mat.df, sum)
  y <- lapply(mat.df, function(x, y) sum(x) + y, y = 5)
  lapply(1:5, function(i) print(i)) # loop
- **sapply(list, function, ..., simplify))** Прилага function върху вектор, матрица или лист. Връща вектор, data.frame, list. Когато simplify=F връща резултат като lapply.
  y2 <- sapply(mat1.df, function(x, y) sum(x) + y, y = 5)
- **tapply(array, indicies, function, ..., simplify))** Прилага function върху елемент от array според групирането от indicies. Връща вектор

или матрица с размери според indicies. Когато simplify=F връща list.
x1 <- runif(16)
cat1 <- rep(1:4, 4)
cat2 <- c(rep(1, 8), rep(2, 8))
mat2.df <- data.frame(x1)
mat2.df$cat1 <- cat1
mat2.df$cat2 <- cat2
tapply(mat2.df$x1, mat2.df$cat1, mean)

3 Агрегация на данни

Агрегирането на данни в R от data.frame става сравнително лесно чрез функцията aggregate(). Тя разделя по подгрупи избраните с by като прилага избраната функция FUN.
dates <- data.frame(date = as.Date("2001-01-01",
format = "%Y − %m − %d") + 0:729, data=0:729)
last.day <- aggregate(x = dates["date"], by = list(month = substr(dates$date, 1, 7)), FUN = max)
last.data <- aggregate(x = dates["data"], by = list(month = substr(dates$date, 1, 7)), FUN = max)

4 Упражнения:

- **Зад. 1.** : Да се свалят данните от Daily share prices (time series).Да се пресметнат средно и дисперсия според обема на търгувани акции.

- **Зад. 2.**: Център на маса (ЦМ) R на система от частици, където всяка е отдалечена от началото на координатната система r(x,y,z) е равна на:

$$R = \frac{\sum m_i r_i}{\sum_i m_i}, \tag{1}$$

  Да се пресметне ЦМ на три частици, с координати A(0,1.5,2), B(2,0,1.5), C(2,1,0.5) и със съответни маси от 2.5, 2.5 и 5 грама.За целта не трябва да се използват цикли.

- **Зад. 3.**: Нека е дадено дискретното съвместно разпределение на две случайни величини X и Y с показаното съвместно разпределение в таблица (1). Да се пресметнат без използването на цикли маргиналните разпределения $P\{X = x_i\}$ и $P\{Y = y_j\}$. Да се направи и проверка за вероятностите.

- **Зад. 4.** Агрегирайте mtcars по cyl и gear и върнете средните стойности.

Table 1:

	0	1	2	3	$P\{X = x_i\}$
0	1/8	1/8	0	0	
1	0	2/8	2/8	0	
2	0	0	1/8	1/8	
$P\{Y = y_j\}$					

- **Зад. 5.** Съберете в обща таблица данните от таблици 'merge1.csv' и 'merge2.csv', като групирате по country и type за:
  - outer joint
  - left joint
  - inner join