BUS41204 Review Session 1

Introduction to R

Jingyu He
jingyuhe@chicagobooth.edu

01/07/2017

Plan

- R, Rstudio
- rmarkdown
- How to do kNN and cross validation in R

What's R

An open-source language for statistical computing and graphics.

- It's popular, have lots of existing packages.
- User-friendly.
- Free!

Someone might ask : why not Python?

- Feel free to use any language you want. Python also has machine learning libraries.
- We recommend R. But we do not provide tech support for any language other than R.

R and Rstudio

- The original R interface is only a command line.
- Rstudio is a fancy interface of R (also free!).
  - See your workspace (variables, functions & data)
  - Write scripts
  - Manipulate files, manage plots
  - It helps a lot when you write a big project.
- You may use Rstudio on your own laptop or Booth's clusters.

Install R and Rstudio

- Download R: `http://cran.r-project.org`
- Download Rstudio: `http://www.rstudio.com/products/rstudio/download`
- Booth online Rstudio: `http://rstudio.chicagobooth.edu`
  - Do computation on remote cluster.
  - Must login with your Booth ID (NOT CNetID)
  - If you don't have a Booth ID, request a temporary one from `helpdesk@chicagobooth.edu`

R packages

R can do many statistical analysis. Functions are organized in “pacakges” or “libraries”. People can make contributions to the community by wrapping their own code as packages.

Install the package

```r
install.packages("package_name")
```

Load the package

```r
library("package_name")
```

Necessary R packages for this course

Run the code below in your R. It takes about 10 minutes to install everything.

```r
packageNames = c("MASS", "ISLR", "animation",
"ElemStatLearn", "glmnet", "textir", "nnet",
"methods", "statmod", "stats", "graphics",
"RCurl", "jsonlite", "tools", "utils",
"data.table", "gbm", "ggplot2", "randomForest",
"tree", "class", "kknn", "e1071",
"data.table", "recommenderlab")

for (pkgName in packageNames) {
if (!(pkgName %in% rownames(installed.packages()))) {
  install.packages(pkgName,
  dependencies=TRUE, repos='http://cran.rstudio.com')
}}

update.packages(ask=FALSE)
```

Set Directory

For each project, put everything under one folder.

```
setwd("path to your project folder")
```

Calculator

Let's begin with some basic examples.

```
10 ^ 2 + 3 - sqrt(10)
```

```
## [1] 99.83772
```

Assign variables

```
a = 3
a^2
```

```
## [1] 9
```

Vector

```r
vec = c(3,4,5)
vec
```

```
## [1] 3 4 5
```

```r
vec = 1:5
vec
```

```
## [1] 1 2 3 4 5
```

```r
vec = seq(0,1,0.1)
vec
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

Matrix

For example, we define a 2-by-3 matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

```
mat = c(1,2,3,4,5,6)
mat = matrix(mat, 2, 3, byrow = TRUE)
mat
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

Transpose a matrix

```
t(mat)
```

```
##      [,1] [,2]
## [1,]    1    4
## [2,]    2    5
## [3,]    3    6
```

Sub-matrix

First row

```
mat[1,]
```

```
## [1] 1 2 3
```

Second column

```
mat[,2]
```

```
## [1] 2 5
```

```
mat[1,2]
```

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

Matrix Mutiplication

Calculate $A \times A^T$. %*% is operator for matrix mutiplication.

```
mat %*% t(mat)
```

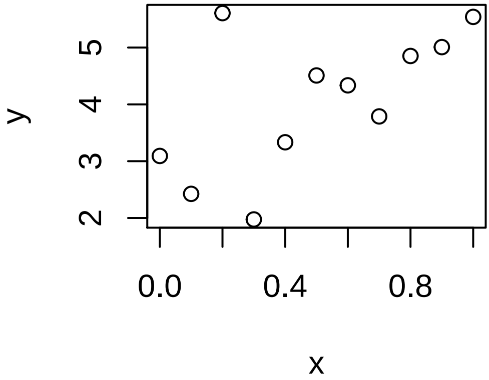
```
##      [,1] [,2]
## [1,]   14   32
## [2,]   32   77
```

```
t(mat) %*% mat
```

```
##      [,1] [,2] [,3]
## [1,]   17   22   27
## [2,]   22   29   36
## [3,]   27   36   45
```

Plot

```
x = seq(0, 1, 0.1)
y = 3 + 2 * x + rnorm(11)
plot(x, y)
```



Data Frame

A matrix with columns name. All variables should have the same length.

```
x = c(11, 12, 14)
y = c(19, 20, 21)
z = c(10, 9, 22)
t = data.frame(x, y, z)
t
```

```
##    x  y  z
## 1 11 19 10
## 2 12 20  9
## 3 14 21 22
```

```
t$x
```

```
## [1] 11 12 14
```

List

List is more general than data frame. Elements of a list can be anything like a matrix, vector, scalar, string or even another list!

```r
xx = c(11, 12, 14)
yy = c(19, 20)
zz = "machine learning"
t = list(x = xx, y = yy, z = zz)
t
```

```
## $x
## [1] 11 12 14
##
## $y
## [1] 19 20
##
## $z
## [1] "machine learning"
```

List

```
t$z
```

```
## [1] "machine learning"
```

Good R tutorials

See the course webpage
`https://chicagoboothml.github.io/ML2016/computing/`

rmarkdown

Generate high quality documents with R raw code and outputs.

- a package "rmarkdown" + Rstudio + latex
- Write in markdown language, the package can compile documents for you.
- This slides was written in this way.
- Writing your homework / project by markdown is strongly recommended.
- Find more info in `https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet`, `http://chicagoboothml.github.io/MachineLearning_Fall2015/Tutorials/R%20Markdown%20Tutorial/`
- A template `https://raw.githubusercontent.com/ChicagoBoothML/ML2016/master/code/BostonHousing_KNN_BiasVarTradeOff_CrossValid.Rmd`

Seeking help

- Read R documents

```r
help(kknn)
help("any function name")
```

- Google it!
- Ask TA, professor or your classmates!
  - Ask questions on piazza. Everyone can see it.