0_MLPS_R_introduction

Zhe Zhang (TA - Heinz CMU PhD)

1/19/2017

Recommended Software

I recommend the use of RStudio program for most of your R coding. Download at https://www.rstudio.com/, which is free for personal use. All packages to use in R will be discussed below or in later cheatsheets.

RStudio is great because it combines several important pieces of your coding environment in one place. I prefer to have my coding on one half of the screen and the "Console" output – where I can debug and test code – on the other half of the screen. Plots, images, and help files then come up occasionally. *This isn't the default.* To adjust this, go to RStudio's Preferences > Pane Layout.

This document is being made in an RMarkdown journal format, which allows easy integration of (1) R code, (2) images, and (3) free text. I *highly suggest* if you're using R for assignments to use RMarkdown journals, which make it easier for both you and the graders. Learn more at http://rmarkdown.rstudio.com/lesson-1.html. Please take a few moments on this helpful tutorial so that you can write your projects and HW effectively. (Also, save this reference sheet: https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf.)

R Basics

This is a brief overview and set of brief tutorials for people who may be coming from another programming language or new to R in general.

I personally am a fan of the work done by the RStudio development team including Hadley Wickham. A great reference for beginning with R for Machine Learning for Problem Solving is Hadley's book, *R for Data Science*, co-authored with Garrett Grolemund. It is available in a easy-to-read online format at: http://r4ds.had.co.nz/introduction.html. I will draw on this and point to examples here and suggest you consult this as a reference.

Furthermore, I recommend starting with R by using several packages that the book also focuses on:

- `ggplot2` package (visualization/visualisation)
  - visualization is a key for any machine learning work. It helps understand the data so that we can get a better sense of how to approach the problem and potential issues to address in the data.
  - visualization also is very important for displaying information to others and for your own understanding in the future
- `dplyr` package (dataset manipulation and transformation)
- in this class, we will mostly use provided rectangular datasets. Each time, you will want to perform several adjustments on these datasets. `dplyr` makes these adjustments with a consistent and simple syntax. Below are the key actions:
  - `summarise` (getting summary statistics)
  - `group_by` (aggregating/summarizing the data differently for particular subgroups of the data)
  - `mutate` (creating new columns of data)
  - `select` (removing or keeping particular columns in the dataset)
  - `arrange` (sorting the rows of the dataset, or sorting by each subgroup)
  - `filter` (keeping or removing certain rows of the data)
  - *many other commands that are occasionally useful and can be found from StackOverflow and the manual, https://cran.r-project.org/web/packages/dplyr/dplyr.pdf*
- `tidyr` package (cleaning and re-arranging the dataset) (more on this later)

- `stringr` package (a clean syntax to deal with strings and text data)

All of these packages are wrapped up into one package, `tidyverse` package. You will see this referred to in the R4DS book.

Key R Commands to Know

This is a living document that will continue to update. For now, I will link to specific chapters in the R for Data Science book to ensure appropriate R background. Please contact me for more information. Here's my recommended reading order:

- Overview of R ML programming http://r4ds.had.co.nz/explore-intro.html
- Dealing with objects basics http://r4ds.had.co.nz/workflow-basics.html
- Visualization basics (3.1 - 3.6) http://r4ds.had.co.nz/data-visualisation.html
  - This is quite important to feel comfortable with because it is the important first step of MLPS and helps you feel connected with the data. I emphasize the practice here.
  - **We will cover this in more detail in the `1_R_MLPS_visualization` file.
- Data transformation (all sections) http://r4ds.had.co.nz/transform.html
- Programming R scripts http://r4ds.had.co.nz/workflow-scripts.html

Additionally, as you get more familiar with R, I suggest you look at each of the sections mentioned here and read up on the examples as you begin to want to write functions, for loops, and piping several commands together. I particularly recommend looking pipes as it is not as commonly seen and works very well with the `dplyr` package. * http://r4ds.had.co.nz/program-intro.html * http://r4ds.had.co.nz/pipes.html

Lastly, I anticipate as you start doing a lot more dataset manipulation, perhaps for you project, I encourage a closer look at how `dplyr` functions. * https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html