Data Science Problem set2

salimeh Birang

January 2020

Data Science tools

1 Statistical Programming Languages

There are many programming languages one can use for statistical analysis namely: R, Python, Julia, Stata, SAS, SPSS, Matlab and JavaScript. Many of these programming languages are built on C,C++ or Fortran. Among the above-mentioned programming languages that has been used for data science, Julia is the newest language which is established in 2012 therefore, it has the smallest community of users compared to R and Python that are established respectively, 1995 and 1991. Each has its own advantages and disadvantages, for instance Julia has the highest speed compared o the other two however, due to being new it is still developing so it is exposed to many changes every year the code you used before might not work later on. Among these languages, R is more specific to data science and very popular tool in this field.

2 Web Scraping tools

One of the tools of data science is the ability to use the data that is being collected constantly through internet.

2.1 How does web scraping work?

Web Scraping involves one of the two tasks either uses an application program interface(API) to download data or downloads HTML files parse the text to extract data.
Many companies like Twitter, Facebook, Linkedin, ... use APIs to guard their data and limit the information can be scraped. Therefore, in these cases it would be better to download HTML code from website and analyze the data. Also it is the only option to extract data for the websites that don't have APIs. However, this approach has its limitation too, if you try to ping too frequently the websites that track the IP addresses of its visitors, you can be blocked.

2.2 web scraping in data science languages

Any programming languages can be used for web scraping, however, it would be easier to use either R, Python or Julia since they have built in packages to parse HTML blocks and load data into tabular environment.

3 Big Data management software

Sometimes you might have huge data set that your computer is not capable of handling it. In order to work with the data set you might need to split it into manageable chunks. But this might not be the best solution if data set gets updated or you want to get summary statistics on the full set of data and so forth.

3.1 RDDs

Resilient Distributed Datasets(RDDs) can alleviate this problem. you'll need a cluster of computers and Hadoop or Spark softwares. Spark cuts down the huge dataset into manageable chunks and run actions on those chunks in parallel, the advantage of using Spark is that in case one of the machine on the cluster fail its data will be transferred to another machine.

3.2 SQL

while having the characteristics of handling huge datasets, SQL can be used to transform data into a form that can be easily process in statistical softwares.

4 Visualization

4.1 ggplot2 (R)

ggplot2 is visualization package in R tidyverse.

4.2 matplotlib (Python)

matplotlib is the graphic package for Python. It is similar to Matlab's visualization syntax.

4.3 plots.jl (Julia)

plots.jl is visualization package for Julia, and it can use any other package like ggplot2 to plot the graph. it enables the user to have different styles.

4.4 Tableau

Tableau is commonly used visualization tool and it build itself as interactive product.

5 Modeling

After collecting data, cleaning it and visualizing it using tools explained we can do some statistical modeling.
1. we can use data to test theories.
2. we can use data to predict behavior.
3. we can use data to explain behavior.