Open Source in Statistical Computation

Moorthy

Open Source in Statistical Computation

Created by Moorthy / @mskmoorthy and modified by Wesley Turner / @wdturner

Learning Objectives

1. Become familiar with the breadth of Open Source Code used in Statistical Computation
2. Gain some practical experience with Open Source alternatives to closed source code
3. Learn how Statistical Computing principles are used in practice for data analysis

Open Source in Statistical Computation - What and Why Now

Statistics is an age old discipline - 17th century according to Wikipedia

Statisticians study data collection, planning of experiments, organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions.

Uses: Probability, sampling, measurement, estimation, least squares, clustering, regression, and design of experiments

Data Science

- Abundance of Data - Video, Image, PDF, Blog, and Newspapers
- Need to Analyze the data - Data Mining, Data Analytics, Data Prediction
- Computational paradigms such as Map-Reduce (Hadoop - Apache 2.0), cloud storage, and data storage technology make handling and analyzing the data possible
- Some computational principles involve Statistical Computing (Monte Carlo)

Computational Languages for Statistical Computing

- SPSS (Proprietary with some open source extensions)
- SAS (Proprietary)
- SciPy (Primarily BSD 3-Clause), Panda (BSD)
- Mondrian (Eclipse v1)
- Shogun (Primarily BSD 3-Clause)
- Perl Data Language (Same as Perl i.e. GPL or Artistic)
- R (Primarily GPL 2 - but see here), RStudio (AGPL v3)

Power of R and RStudio

- Open Source Software
- A lot of Mathematical and Statistical packages available
- Installing is easy
- Integrated IDE
- Literate Computing - See Shiny
- Plotting Packages
- A large number of Data Sets are available with R

The Power of Communities

- Conferences
- Journals

Difficulties with R

- New programming language and syntax
- All Data are stored in memory
- New Strategy has to be adopted if the data does not fit in memory
- R is not a general purpose programming language

Data Science Questions - Level 1

- Develop Expectations
- Collect Data
- Match Expectations with Data

Data Science Questions - Level 2

- Stating the Question
- Exploratory Data Analytics
- Model Building
- Interpret
- Communicate

Caveat

- Data can be used to answer many questions but not all of them

Tuckey's Quotation

- The data may not contain the answer.
- The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

Download

- Download R and RStudio
- Download http://rtutorialseries.blogspot.com/

Warm up exercise

We will do all the examples in Introduction to R, Descriptive Statistics, and Data Visualization

The End

by Moorthy