STA 141A
Fundamentals of Statistical Data Science

Fall 2016

Instructor: Debashis Paul

Logistics

- **Lecture: TR 9:00 – 10:20 AM (Everson 176)**
- **Discussion:**
  - **F 9:00 – 9:50 AM (Olson 206)**
  - **F 10:00 – 10:50 AM (Olson 206)**
- **Instructor Office Hours:**
  - **T 1:00 – 2:30 PM (MSB 4208)**
  - **R 10:30 AM – 12:00 PM (MSB 1143)**
- **Instructor Email:** debpaul@ucdavis.edu

Teaching Assistants

- **Haoran Li (hrli@ucdavis.edu)**
- **Nicholas Ulle (naulle@ucdavis.edu)**
- **Aoran Zhang (arzhang@ucdavis.edu)**
- Discussion sections will be led by Nick Ulle
- Office hours of the TAs will be announced soon

What to learn from this course ?

- Handling data of different kinds, including data in irregular formats
- Core statistical principles for summarizing complex data
- Data visualization techniques
- Developing algorithms for data analysis and implementing them in R programming language
- Techniques and principles for statistical simulation
- Basics of statistical learning theory and practice
- Preparing reports and documenting software

Follow up courses

- STA 141B : *Data and Web Technologies for Data Analysis*

1. Essentials of using relational databases and SQL
2. Scraping Web pages and using Web services/APIs
3. Basics of text mining
4. Interactive data visualization with Web technologies
5. Computational data workflow and best practices
6. Statistical machine learning methods
7. Will use Python programming language

Follow up courses

- STA 141C : *Big Data and High Performance Statistical Computing*

1. High-performance computing
2. Distributed and parallel computing, algorithm and computational reasoning
3. Different computational approaches and paradigms for analysis of big data
4. Interfaces to compiled languages
5. Will use Python programming language

Follow up courses

- STA 160 : *Practice in Statistical Data Science*

This course serves as a capstone course in which the students focus on the practice of data analysis, and both statistical and computational reasoning. Students will work in groups on a data analysis project with the following emphasis:

(a) frame the question and possible approaches
(b) acquire data (if necessary)
(c) clean and explore the data
(d) use appropriate statistical and machine learning methods to effectively answer the question(s)
(e) prepare a technical report & presentation (for a non-statistical audience) detailing the conclusions and insights, potential shortcomings/issues, and possible alternative approaches and directions.

Introduction to R

- R home : https://cran.r-project.org/
- Download and install the precompiled binary distributions of the base system. Versions for Linux, Windows and Mac OS X are available.
- CRAN website has plenty of resources, including an excellent FAQ page
- Manuals on *An Introduction to* R, R *Data Import/Export*, R *Installation and Administration*, *Writing* R *Extensions*, and *The* R *Reference Index* are available on https://cran.r-project.org/manuals.html

Vectors

```
x = c(1,2,4,2.5,-1.7,1)  # creates a vector of length 6
class(x)  # type of the "object" x;  returns "numric"
y = x^2
z = sin(x) + cos(x^1.5) * exp(-x^3)
# Both y and z have the same length as x; These are functions of the variable x
x/y  # elementwise division
x %o% y   # outer product of x and y
```

Matrices

M = matrix(0,2,3) # matrix with 2 rows and 3 columns with all entries 0

Xmat1 = matrix(x,2,3) # 2 x 3 matrix formed by elements of vector x,

ordered in column-wise manner

Xmat2 = matrix(x,2,3,byrow=T) # now elements of x are ordered row-wise

t(Xmat1) # transpose of Xmat1

Xmat1 %*% t(Xmat2) # multiplication of 2 x 3 matrix with a 3 x 2 matrix

as.matrix(x) # vector x treated as a 6 x 1 matrix

Arrays

?array # find help on the object “array”

Output:

array(data=NA, dim=length(data), dimnames = NULL)

data : a vector (including a list or expression vector) giving data to fill the array

dim : the dim attribute for the array to be created, that is an integer vector of length one or more giving the maximal indices in each dimension

dimnames : either NULL or the names for the dimensions. This must a list (or it will be ignored) with one component for each dimension, either NULL or a character vector of the length given by dim for that dimension

Basic plotting functions in R

- plot( ) # draw a scatterplot
- points( ) # adding points to an existing plot
- pairs( ) # draw a scatter plot matrix (for 2 or more variables)
- lines( ) # joining points (based on criteria) with straight line segments
- matplot( ) # plotting columns of a matrix against a variable
- abline( ) # draw a single straight line with specified intercept and slope

Basic graphical statistical summaries

- hist( ) # draw histogram (of numeric data)
- boxplot( ) # draw Boxplot (visual description of five-point summary)
- dotchart( ) # draw Dot chart (of numeric data)
- barplot( ) # draw Bar plot
- pie( ) # draw Pie chart (of categorical data)
- qqplot( ), qqline( ), qqnorm( ) # draw Q-Q (quantile vs quantile) plot

Customizing plots

- par( ) # used to set graphical parameters
- axis( ) # customizes axes of a plot
- legend( ) # add a legend to a plot
- text( ), mtext( ) # add text to a plot
- title( ) # add a title to a figure
- box( ) # draw a box around a current plot
- rectangle( ), polygon( ), # draw a rectangle, or a polygon