Statistics 471/571/701 Fall 2016

Modern Data Mining

Mondays and Wednesdays

Session 401: **10:30 – 12:00, JMHH-F55**

Session 402: **3:00 – 4:30pm, JMHH-G55**

Linda Zhao: lzhao@wharton.upenn.edu, Office: JMHH470.

Office Hours: 3:00 to 5:00 pm, Fri or by appointment

Course Description: Statistics has been evolving rapidly to keep up with the modern world. We will show how to modify and adapt simpler models; then go beyond with relatively newer methods/techniques to handle contemporary large and complex data with applications in finance, marketing, medical fields, social science, entertaining… you name it.

This class is cross-listed as STAT471 for undergraduates, STAT571 as a graduate level course for students outside of the statistics department and STAT701 for MBA's.

A brief short list of methods include: Multiple Regression/Stepwise Regression, Logistic Regression, KNN (K nearest neighbor), LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), LASSO, Ridge Regression, PCA, Tree based methods such as Random Forest, Support Vector Machines. Some text mining methods will be introduced. Bootstrap and k-fold cross validation will be used and criterions such as Training and Testing errors, ROC/AUC and FDR are used. Unsupervised learning or other topic will be covered as well. The free but powerful software "R" will be our tool. At the end of the semester we hope that students not only learn the modern statistical methods but also become skilled in dealing with data of essentially any size.

Prerequisites: Two semesters of statistical courses, no programming experience required. Familiarity of multiple regressions is assumed. A quick review about Chapter 3 from the main textbook would be very helpful.

Case studies: Who tweets for Trump? Can we do something to reduce crime rates? Framingham heart disease study; Billion dollar Billy Beane; What can we do to improve education – Texas third graders? Whose political bill is more likely to be approved in the sea of bills proposed by politicians? Can you predict housing prices? McGill Billboard – how long a song can sit on the board? Out of 502 stocks can we do better than S&P500? How to be successful at Kickstarter, a popular crowd fundraiser? Hunting for important gene express positions to help out with HIV+ patients; From Yelp reviews to predict the rating and more…

Computer package: The statistical computing language R is used. There are infinitely many new packages available for us to use. It is open source and it is free.

A friendly user interface of R called Rstudio will be used. Download and install Rstudio: https://www.rstudio.com/products/rstudio/download/

General information about R is available through www.R-project.org

knitr: A dynamic report system which imbedding the R output into our text report. It is very convenient for reproducibility.

R tutorial: 4:00-5:30, TH, Sept. 1 and Friday, Sept. 2 (Bring your laptop)

Two tutorial sessions will be given by TAs.

A laptop is a must for the course. You are encouraged to bring the laptop to the classes so that you may run the lecture code simultaneously with the professor.

Textbook: (Required and they are all available online for free)

1. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani,
An Introduction to Statistical Learning with Application in R (ISLR),
First Edition, 2013, Springer New York.

An e-version is available from the author's website: http://www-bcf.usc.edu/~gareth/ISL/

2. An Introduction to R: http://cran.r-project.org/doc/manuals/R-intro.pdf

Useful Links:

Quick R: http://www.statmethods.net
Popular R packages: https://www.rstudio.com/products/rpackages/

Additional optional reading:

3. Peter Dalgaard,
Introductory Statistics with R,
Second Edition, 2008, Springer

Available as a pdf
http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/Introductory_Statistics_with_R__2nd_ed.pdf

4. Trever Hastie, Robert Tibshirani, Jerome Friedman
The Elements of Statistical Learning: Data Mining, Inference, and Prediction (ESL)
Second Edition, 2008, Springer

A pdf version is available:
http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

Canvas: https://canvas.upenn.edu

Most of the materials including announcements, data, R codes, homework/projects, solutions, etc. will be available on our Canvas site. (Turn on notifications especially for Announcement, files…: Settings > Notifications and go from there.)

I believe that one can only learn by doing. We will put together a few things throughout the semester for you to do. These are:

Homework: 5-**6** structured homework's will be given.

Exams/Projects:

One open book midterm - a laptop is needed: basic questions plus an onsite data analysis with R

A take home mini project (individual): A prelude/preparation for our final project. I will bring you a data set that you go through in the process of data analyses. A report is needed that includes

i) Goal of the study and the findings
ii) Summary of the data
iii) Detailed analyses
iv) R-code

Four very short (10 minutes each but 40 minutes for the last one) in class quizzes: simple multiple-choice questions. The lowest grade among the first three will be dropped.

Final project: The ultimate goal of the class is to prepare/expose students to techniques that are suitable for modern data. A final project is designed so that each of you will bring a problem of your interest to the class. You will need to identify a problem to tackle with a data set that either you collect/extract or find. A complete write up is required. This would be a good project to be put in your CV if needed.

i) Proposal can be submitted any time during the semester but no later than **11/20/Sun**
ii) A well-motivated, relevant topic is most desirable.
iii) Originality, Complexity with challenge will be another plus
iii) A complete write up is a must.

Group work: The homework and the final project can be done by groups of up to three people. Please try to form a group and send the information to us no later than **09/18/Sun**. We will help out for those who need to find a group.

Schedules: All the following exams/projects are individual ones except for the final projects. All the exams are open book.

Midterm: **11/01/Tu**: **6:00-8:00pm**
Take Home mini project: TBA
Four Quizzes:
Quiz 1: TBA
Quiz 2: TBA
Quiz 3: TBA
Quiz 4: TBA

Final Project: 12/18/Sun, 2016

Grade allocation:

Homework: 15%
Quizzes: 15% (The lowest quiz of the first three will be dropped)
Take home min project: 25%
Midterm: 25%
Final Project: 20%
(Professor will make adjustment for those who actively contribute to the class throughout the semester.)

TAs:

Sam Pimentel, sjp@wharton.upenn.edu
Junhui Cai, junhui@seas.upenn.edu
Chris Hua, chua@wharton.upenn.edu
Stacey Sloate, ssloate@sas.upenn.edu
Sam Yarosh, samyarosh@gmail.com