Statistics 471/701 Fall 2015

Modern Data Mining

M, W 10:30 – 12:00 JMHH F60

Linda Zhao: lzhao@wharton.upenn.edu, Office: JMHH470.
Office Hours: 3:00-5:30 Thursdays or by appointment

TA: TBA

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.

A brief short list of methods include: Multiple Regression, Logistic Regression, KNN (K nearest neighbor), LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), LASSO, Ridge Regression, Tree based methods such as Random Forest, Support Vector Machines. Would like to cover some text mining methods. Bootstrap and k-fold cross validation will be used and criterions such as Training and Testing errors, ROC/AUC and FDR are used. 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. (my wish!)

Collection of data: 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 – might be too ambitious for this data because of its un-manageable size…

Computer package: The statistical computing language R will be used. There are infinitely many new packages available for us to use. It is open source and it is free. It is available through www.R-project.org for all common computing platforms such as Windows, Mac and Linux.

R tutorial: TBA at Wharton Computer Lab: JHMM375.

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

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/download.html

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.

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: 4-5 structured homeworks will be given. The lowest grade will be dropped.

Exams/Projects:

One in class, 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

Three very short (10 minutes) in class quizzes: simple multiple-choice questions. The lowest grade 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.
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. 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: Monday, Oct 19th, a laptop is needed!
Take Home mini project: Nov. 14th.
Three Quizzes:
Quiz 1: 09/14/Mon
Quiz 2: 10/05/Mon
Quiz 3: 11/09/Mon

Final Project: Due, Sunday, Dec 20th.

Grade allocation:
Homework: 25% (The lowest grade will be dropped)
Quizzes and take home min project: 20% (The lowest quiz grade will be dropped)
Midterm: 25%
Final Project: 30%
(Professor will make adjustment for those who actively contribute to the class throughout the semester.)

Topics covered:
Multiple Regressions
Logistic Regressions
Model Selections: Cp, BIC ; Misclassification Errors
ROC and AUC; FDR,
LDA, QDA and KNN
Lasso and Ridge Regression
Bootstrap and Cross Validations
Trees, Bagging and Random Forest
Support Vector Machines
Text Mining

Schedules: A tentative schedule, some adjustment maybe needed as semester proceeds.

Lecture	Date	Contents	Note
1	08/26/Wed	Introduction Ch 1-2, **2 lects**	
2	08/31/Mon		
3	09/02/Wed	Simple Regression, Ch 3.1 **2 lects**	
	09/07/Mon	Labor Day	
4	09/09/Wed		
5	09/14/Mon	Multiple Regression, Ch 3.2 – 3.6 **4 lects**	**Quiz 1**
6	09/16/Wed		
7	09/21/Mon		
8	09/23/Wed		
9	09/28/Mon	Classifications, logistic reg, Ch 4.1 – 4.3 **3 lects**	
10	09/30/Wed		
	01/02/Fri		
11	10/05/Mon		**Quiz 2**
12	10/07/Wed	Classifications, LDA/QDA, 4.4-4.6 **3 lects**	
	10/08/Th	Fall Break	
13	10/12/Mon		
14	10/14/Wed		
15	10/19/Mon	Midterm (in class)	**Midterm**
16	10/21/Wed	Resampling, Ch 5 **2 lects**	
17	10/26/Mon		
18	10/28/Wed	Model selection, Ch 6 **5 lects**	
19	11/02/Mon		
20	11/04/Wed		
21	11/09/Mon		**Quiz 3**
22	11/11/Wed		
	11/15/Sun	Take home project due	**Take home Project**
23	11/16/Mon	Text mining (Extra material) **3 lects**	
24	11/18/Wed		
25	11/23/Mon		
	11/25/Wed	No class	

	11/26/Th	Thanksgiving	
26	11/30/Mon	Tree based method, Cha 8 **3 lects**	
27	12/02/Wed		
28	12/07/Mon		
	12/08/Tu	Last Day	
	12/20/Sun	Before 11:59pm	**Final Project Due**