Statistics 471/571/701: Modern Data Mining

Linda Zhao

Fall 2019

E-mail: lzhao@wharton.upenn.edu
Web: canvas.upenn.edu
Office: JMHH 470
Office Hours: F 3:00 - 5:00 pm, or by appointment
Class Room: JMHH 250
Class Hours: 401: MW 10:30am - 12:00pm; 402: MW 3:00 - 4:20pm; 403: MW: 1:30 - 3:00pm

---

Modern Data Mining

Course Description

Statistics has been evolving rapidly to keep up with the modern world, especially with computational methods for the explosion of data. As a signinificant part of data science we start the class with exploratory data analysis (EDA). We then show how to build, interpret, and adapt simple models; then go beyond with newer contemporary methods and techniques for handling large and complex data with applications in finance, marketing, medical fields, social science, entertainment, you name it. While this course makes extensive use of the statistical programming language R, no programming experience is required. By the end of the semester we hope that students have not only learned the modern statistical methods but have also become skilled in dealing with data of essentially any size.

This class is cross-listed as STAT 471 for undergraduates, STAT 571 as a graduate level course for students outside of the statistics department, and STAT 701 for MBA's.

Prerequisites

Two semesters of statistics courses, familiarity with multiple regressions is assumed. A quick review of Chapter 3 from the main textbook (ISLR) would be very helpful.

Methods covered (mostly)

- Exploratory Data Analysis (EDA)
- Multiple Regression/Stepwise Regression
- Logistic Regression/Multi-Nomial regression
- K-nearest neighbors (KNN)
- Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA)

- Penalized regression: LASSO, Ridge Regression, Elastic Net
- Text mining sentiment analysis
- Neural network/Deep learning
- Tree based methods such as Boosting, Random Forest
- Support Vector Machines.
- Bootstrap and k-fold cross validation
- Training and Testing errors
- ROC/AUC and FDR
- Principal Components Analysis (PCA)
- Network model
- Unsupervised learning

Case study/Datasets

Most of the following cases will be covered:

- Who tweets for Trump?
- Wharton Business Radio Audience Estimation via Amazon Mturk
- Diabetes/Healthcare (Predicting Readmission Probability for Diabetes Inpatients to Save Healthcare Cost)
- IQ=Success?
- Boost return by 80% in Lending Club?
- Handwriting recognition (image recognition)
- 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
- Hunting for important gene expression positions to help out with HIV positive patients
- Using Yelp reviews to predict the rating (text minging)
- Chinese Annual Industrial Survey
- Gene expression data miracles
- Seattle housing price

And more!

Course Materials

Software

The free and open source statistical computing language R is used. There are infinitely many new packages available for us to use; a pretty interface to explore the publicly available R packages is available via Microsoft.

Throughout of the semester, we use the free RStudio, an interface for writing R documents and working with data.

R tutorial (Optional: Check canvas for locations)

TA's will run a basic R tutorial during the first week of class. Bring a laptop.

- 08/27 (Tue): 4:30 - 5:30 PM Location TBA

An advanced R tutorial convering `dplyr`, `data.table` and `ggplot` will be given at (two same sessions.)

- 08/30 (Fri): 4:30 - 6:00 PM Location TBA
- 09/04 (Wed): 4:30 - 6:00 PM Location TBA

Textbook

Our required textbook is known as ISLR and is freely available from the authors:

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

Additional optional readings are recommended for getting familiar with and working with R. These include:

- Garrett Grolemund & Hadley Wickham, *R for Data Science*, 2016, O'Reilly. Available freely online.
- The R Core Team, *An Introduction to R*, available from CRAN.
- Patrick Burns, *The R Inferno*, available online.
- Peter Dalgaard, Introductory Statistics with R, Second Edition, 2008, Springer. Available on Academia.

An advanced text book as a reference:

- Trever Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, 2008, Springer. A PDF version of ESLR is available from the authors.

A reference for general statistics method you may read:

- Ramsey and Schafer, *The Statistical Sleuth*, Third Edition, 2013, Brooks/Cole (an e-version is available in the canvas site)

Course Policies

Communication

Communication will be through Canvas and through Piazza. Files will be uploaded to Canvas, including datasets, homeworks, and lecture notes. Piazza is a useful forum for students to ask questions.

Laptop Policy

A **laptop** is a must for the course. You are encouraged to bring the laptop to classes so that you may run the lecture code simultaneously with the professor. However, it is not allowed to use the laptop for other purposes during the lectures. Cell phones must be turned off.

Assignments and Exams

Homework: We will give 4 or 5 homework assignments. These may be done in groups of up to 3 people; see the Group Policy for more details.

Mini project (Sun 11/17): This will be a take-home, *individual*, assignment. You will be given a dataset to analyze, step by step. There is a limit of **10 pages**. Further information will be given.

Quizzes: We will give 4 short in-class individual quizzes. The first three will be 10-minute each, and the lowest grade among these will be dropped. The final quiz will be 40-minute and cannot be dropped. *No makeup quizzes will be offered.*

1. **Quiz 1: Mon 9/16**
2. **Quiz 2: Mon 10/7**
3. **Quiz 3: Mon 10/21**
4. **Quiz 4: Mon 12/9**

Midterm: Tuesday, 11/5, 6:00-8:00 PM This exam will be *individual*, open-book and done on the computer. You will be given an exam to work through, using R, to do under proctored conditions. You must bring a laptop.

- 401: SHDH 350
- 402: SHDH 351
- 403: SHDH 213 and 1206

Final Project (Sun 12/15): The ultimate goal of the class is to prepare/expose students to techniques that are suitable for modern data. The final project is designed so that each of you will bring a problem of personal interest to the class. You will need to identify a problem to tackle with a data set that either you collect/extract or find.

This project is done within a group of up to 3 members. A complete write up is required. This would be a good project to put in your CV if desired.

- A well-motivated, relevant topic is most desirable.
- Originality, complexity, and challenge will be another plus
- A complete write up is a must.
- **Maximum of 15 pages**.
- Kaggle: is a good place to find a data set.

Data Science Live (DSL) (TBA)

- Showcase your final project to fellow students, invited firms and public
- Peer reviews account for the final project grade
- Check canvas for updates

DSL, Spring 2019

Late Work Policy

It is imperative that you manage your workload properly for this course. We will allow late assignments up to 3 days late, with a 15% penalty per day. Note that lateness will be determined by timestamp on Canvas submissions, i.e. 12:01 AM is considered late.

Note also the above quiz policy: we will drop your lowest score of the first 3 quizzes, but there will not be any makeup quizzes.

Group Policy

The homework and the final project can be done by groups of up to three people. Sign up for groups on Canvas as soon as possible but no later than **Wednesday, 9/11**. We will help out for those who need to find a group, with searches on Piazza.

Please note that at no time may a group have more than 3 members. In addition, while those within a group will submit a single homework file for the group, students must follow the code of academic integrity in regards to classmates outside their group. Finally, students do not have to complete the final project in the same group as for homework. They may form a new group though again no more than 3 people may be in a group. We prefer you keep the same groups through the semester but it is now required.

Grading Policy

- Homework: 20%
- Quizzes: 20%
- Mini Project (take-home): 15%
- Midterm Exam: 25%
- Final Project: 20%

Professor Zhao may make adjustments for those who actively contribute to the class throughout the semester.

Teaching Assistants

- Lead TAs:
  - Arun Kumar Kuchibhotla `arunku@wharton.upenn.edu`
  - Harrison Beard `hbeard@sas.upenn.edu`
- Chenyang Fang `cyfang@wharton.upenn.edu`, Aston Hamilton `hasto@wharton.upenn.edu`, Farnik Nikakhtar `farnik@sas.upenn.edu`, Arielle Stern `arstern@seas.upenn.edu`, Trevor Wexner `wexnert@sas.upenn.edu`

R Markdown Advice

Make sure to include your Rmd file and a knitted file, in your submission.

1. Load your packages only once, and at the top of the document, in the setup chunk.
2. Use `library(pkg_name)` to load packages. Do not use `require()` - see Hadley's advice.

3. You can use `command-option-i` to insert a new R chunk.
4. Most of the time, you will want to use `stringsAsFactors = FALSE` when reading csv files.
5. Do not include `install.packages("...")` in your code.
6. Any time R prints output, make sure to include text explaining what is being printed and why.

Class Schedule

Tentative and subject to change. Unless otherwise noted, readings refer to *Introduction to Statistical Learning*.

Week 01, 08/26 - 08/30:

- **Wed 8/28**: EDA/Simple regression

Week 02, 09/02 - 09/06:

- **Mon 9/2**: Labor Day (no class)
- **Wed 9/4**: EDA/Simple regression

Week 03, 09/09 - 09/13:

- **Mon 9/9**: Multiple regression (Ch 3.2 - 3.6)
- **Wed 9/11**: Multiple regression
- **Wed 9/11**: **Grouping due on Canvas**
- **Sun 9/15**: **Homework 1 due**, before 11:59 PM to Canvas

Week 04, 09/16 - 09/20:

- **Mon 9/16**: **Quiz 1**. Continued topics
- **Wed 9/18**: Model selection (Ch 6.1)

Week 05, 09/23 - 09/27:

- **Mon 9/23**: LASSO (Ch 6.2)
- **Wed 9/25**: Continued topics

Week 06, 09/30 - 10/04:

- **Mon 9/30**: Logistic regression, MLE (Ch 4.1-4.3)
- **Wed 10/2**: Continued topics
- **Sun 10/6**: **Homework 2 due**, before 11:59 PM to Canvas.

Week 07, 10/07 - 10/11:

- **Mon 10/7**: **Quiz 2** Classification (ROC, AUC, FDR). Bayes rule
- **Wed 10/9**: Continued topics
- **Wed 10/9**: Fall break

Week 08, 10/14 - 10/18:

- **Mon 10/14**: LASSO for classification/Text mining Multi-level Classification
- **Wed 10/16**: Multi-level Classification

Week 09, 10/21 - 10/25:

- **Mon 10/21**: **Quiz 3**, Neural network/deep learning
- **Wed 10/23**: Continued topics
- **Sun 10/27**: **Homework 3 due**, before 11:59 PM to Canvas.

Week 10, 10/28 - 11/01:

- **Mon 10/28**: Leeway
- **Wed 10/30**: Decision trees (Ch 8.1)

Week 11, 11/04 - 11/08:

- **Mon 11/4**: Bootstrap/Bagging (Ch 8.2)
- **Tue 11/5**: **Midterm Exam** 6-8PM, Location: SH-DH350/351/213,1206.
- **Wed 11/6**: Random Forest (Ch 8.2)
- **Wed 11/6**: **Mini project out**

Week 12, 11/11 - 11/15:

- **Mon 11/11**: Random Forest
- **Wed 11/13**: Boosting (Ch 8.2)
- **Sun 11/17**: **Mini project due**, before 11:59 PM to Canvas.

Week 13, 11/18 - 11/22:

- **Mon 11/18**: Leeway
- **Wed 11/20**: Principal component analysis (PCA)

Week 14, 11/25 - 11/29:

- **Mon 11/25**: Principal component analysis (PCA)/Clustering
- **Wed 11/27**: No class (Thanksgiving)
- **Sun 12/1**: **Homework 4 due**, before 11:59 PM to Canvas.

Week 15, 12/02 - 12/06:

- **Mon 12/2**: Continued topics
- **Wed 12/4**: Continued topics
- **Sun 12/8**: **Final project due**, before 11:59 PM to Canvas

Week 16, 12/09 - 12/13:

- **Mon 12/9 (Last Class): Quiz 4**. Last class: Final quiz