Statistics 571/701: Modern Data Mining

Linda Zhao, Spring 2021

E-mail: lzhao@wharton.upenn.edu
Web: piazza.com/upenn/spring2021/stat571stat701
Office: WARB 403 (Wharton Academic Research Building (new))
Office Hours: 3:00-5:00pm Fridays or by appointment
Class Room: TBA
Class Hours: TH: 1:30 - 3:00pm

Modern Data Mining

Course Description

Statistics has been evolving rapidly in the era of big data and provides tools to harvest knowledge from big data. Focusing on methodologies with reasoning, the class brings in a large set of cutting edge machine learning techniques with applications. Hands-on data experience with R throughout the semester is another feature. We divide the course into three portions: supervised learning, non-supervised learning, a well-motivated self-designed final project. The class will begin with data acquisition and exploratory data analysis (EDA) along with tools for reproducible report, an essential part of data science. We next show how to build, interpret, and adopt basic models; then go beyond to 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 extensively uses the statistical programming language R, no programming experience is required. By the end of the semester, students will master popular modern statistical methods but also get equipped with hands-on skills in handling data of essentially any size.

This class is cross-listed as STAT 571 for graduate students, and STAT 701 for MBA's. **Permission needed for advanced undergraduate students**. Please fill the Google form (https://forms.gle/bkpNTCsf3tCBVLMi8).

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)

- R/Rstudio/Knitr
- Exploratory Data Analysis (EDA)
- Multiple Regression
- Robust standard error estimation
- Stepwise Regression (Cp/AIC, BIC)
- Training and Testing errors
- k-fold cross validation
- Penalized regression: LASSO, Ridge Regression, Elastic Net
- Logistic Regression/Multi-Nomial regression
- Classification/ROC/AUC and FDR
- Bootstrap
- Tree based methods (Bagging, Random Forest and Boosting)
- Neural network/Deep learning
- Text mining sentiment analysis
- Principal Components Analysis (PCA)
- Unsupervised learning
- Network model
- K-nearest neighbors (KNN)
- Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA)
- Support Vector Machines
- Missing Data

Case study/Datasets

Most of the following cases will be covered:

- Lockdown/Compliences/Covid19
- 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 mining)
- 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 through Rstudio. 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.

- 01/20 (Wed): 4:30 - 5:30 PM Location TBA

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

- 01/22 (Fri): 4:30 - 6:00 PM Location TBA
- 01/25 (Mon): 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.

Quizzes: We will give two 10-minute in-class individual quizzes. The final quiz will be 40-minute as a final exam. *No makeup quizzes will be offered.* Contact the instructor in advance for special cases.

1. **Quiz 1: 2/9 (Tue)**
2. **Quiz 2: 3/2 (Tue)**
3. **Quiz 3: 4/29 (Thu)**

Midterm: 3/29 (Mon), 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.

Final Project (Sun 5/2): 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) (Friday, 4/30)

- 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
- DSL, Fall 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.

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 **Thursday, 1/28**. 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: 30%
- Quizzes: 20% (5% for quiz 1 & 2; 10% for quiz 3)
- Midterm Exam: 30%
- Final Project: 20%

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

Teaching Assistants

TBA

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.