STAT 471/571/701: Modern Data Mining

Linda Zhao, Spring 2022

E-mail: lzhao@wharton.upenn.edu
Office: WARB 403
Class Room: JMHH 240

Web: piazza.com/wharton.upenn/spring2022/stat471571701
Office Hours: 4:00-6:00 pm Fridays or by appointment
Class Hours: TR: 10:15 - 11:45am (401); 12:00 - 1:30pm (402)

Modern Data Mining

Course Description

Statistical/Machine Learning has been evolving rapidly in the era of big data and provides tools for scientific discoveries and data-driven decision-making. Focusing on methodologies with statistical reasoning, the course brings in a large set of cutting-edge machine learning techniques combined with up-to-date case studies. Hands-on experience with R throughout the semester is another feature. The course covers data science essentials starting from data acquisition and exploratory data analysis (EDA) along with tools for reproducible reports (Rmarkdown). We next show how to build and interpret basic models; then we go beyond and focus on 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.

Lecture notes will be provided by the instructor. They are organized by topics (modules) and written in the reproducible RMarkdown format which combines description, visualizations, explanation, and R codes. This class is cross-listed as STAT 471 for undergraduates, STAT 571 for graduate students, and STAT 701 for MBAs.

Prerequisites

Two semesters of statistics courses, familiarity with multiple regressions is assumed. To prepare yourself with the data science workflow, R for Data Science is a good reference.

Methods covered (mostly)

Part I: Acquiring, preparing, exploring and visualizing data

- R/Rstudio/Knitr
- Study design and data acquisition/preparation
- Exploratory Data Analysis (EDA)
- Principal Components Analysis (PCA)
- Clustering
- Missing data

Part II: Model-based supervised learning

- Multiple regression
- Robust standard error estimation
- AB testing/Multiple testing
- Step-wise regression (Cp/AIC, BIC)
- Training and testing errors
- k-fold cross validation
- Bootstrap
- Penalized regression: LASSO, Ridge Regression, Elastic Net
- Logistic Regression/Multi-Nomial regression
- Classification/ROC/AUC and FDR
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

Part III: Machine learning

- K-nearest neighbors (KNN)
- Tree based methods (Bagging, Random Forest and Boosting)
- Support Vector Machines (SVM)
- Neural network/Deep learning
- Text mining/Nature Language Processing (NLP)
- Network model
- Matrix completion (Recommendation system)

Case study/Datasets

Most of the following cases will be covered:

- COVID-19: Lock-down and Compliance
- Who tweets for Trump?
- Wharton Business Radio Audience Estimation via Amazon Mturk
- Diabetes/Health care (Predicting Readmission Probability for Diabetes Inpatients to Save Health care 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. We will use RMarkdown for all materials to ensue reproducibility. We will also use Git and GitHub for version control and collaboration.

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

Install the following software: R, RStudio, RMarkdown and git. Detailed instructions are available in canvas.

R tutorial

- Basic R tutorial (Available in canvas)
  - Get_staRted.Rmd/Get_staRted.html
  - Video recording
- An advanced R tutorial (First lecture in class on Jan. 13th)
  - advanced_R_tutorial.Rmd
  - covering `dplyr`, `data.table` and `ggplot`

Lecture notes

Over the years we have been developing our own lecture notes. They are organized by topic and written in reproducible RMarkdown format which combines R codes, visualizations, and narrative text. Real case studies are deployed throughout. The methods are explained with insightful ideas with minimum mathematics. Some deeper explanations and useful materials are postponed in Appendices as references. Students are urged to read through before classes and put hands on line by line at some point.

Textbooks

While we suggest you to read through thoroughly our own lecture notes which often cover more materials we require you to have the following two books:

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Application in R (ISLR)*, Available freely online, Second Edition, 2021, Springer New York.
- Garrett Grolemund & Hadley Wickham, *R for Data Science*, 2016, O'Reilly. Available freely online.

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

- The R Core Team, *An Introduction to R*, available from CRAN.
- Hadley Wickham, *Advanced R*, 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/answer 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 can 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. The higher grade of the two short quizzes will be doubled.* Contact the instructor in advance for special cases.

1. **Quiz 1: 2/8 (Tue)**

2. **Quiz 2: 3/1 (Tue)**
3. **Quiz 3: 4/26 (Tue)**

Midterm: 3/28 (Mon), 7:00-9:00 PM This exam will be an in class, *individual*, open-book and done on the computer. You will be given an exam in RMarkdown format to work through. All TAs will be available to answer questions. Previous exams will be available on Canvas.

Final Project: 5/1 (Sun): 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, collect/extract or find an appropriate data set, run a complete data driven study and make a final conclusion from your study.

This project is done with 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.
- Setup a GitHub Page for the project.
- **Maximum of 15 pages**.
- Some data sources:
  - Kaggle is a good place to find a data set.
  - Google provides public dataset through BigQuery on Google Cloud Platform.
  - gapminder
  - UCI Machine Learning Repo

Data Science Live (DSL) (Friday, 4/29, 5:00 - 7:00 PM)

We will run a workshop to showcase selected final projects to students campus-wise.

- Groups intested in presenting in DSL can submit proposals any time before the end of **4/22**. We will select outstanding groups to present their projects.
- Selected groups receive **bonus** project grades.
- DSL, Spring 2019
- DSL, Fall 2019
- DSL, Spring 2021

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 the 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 (can be from either sections). Sign up for groups on Canvas as soon as possible but no later than **Tuesday, 1/26**. 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: 15% (6% for quiz 1 & 2; 9% for quiz 3)
- Midterm Exam: 30%
- Final Project: 25%

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

Class Schedule

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

Week 01, 01/10 - 01/16:

- **Thu 1/13**: Advanced R tutorial (dplyr/data.table/ggplot)

Week 02, 01/17 - 01/23:

- **Tue 1/18**: Data acquisition and preparation
- **Thu 1/20**: Exploratory data analysis (EDA)

Week 03, 01/24 - 01/30:

- **Tue 1/25**: Dimension Reduction/Principal Component Analysis (PCA), (Ch 6.3.1, 12.1, 12.2)
- **Tue 1/25**: **Grouping due on Canvas**
- **Thu 1/27**: Clustering (Ch 12.4.1 and 12.4.3)
- **Sun 1/30**: **Homework 1 due**, before 11:59 PM to Canvas

Week 04, 01/31 - 02/06:

- **Tue 2/1**: **Quiz 1**. Linear regression (Ch 3.1 - 3.6)
- **Thu 2/3**: Continued topics

Week 05, 02/07 - 02/13:

- **Tue 2/8**: Continued topics
- **Thu 2/10**: Model selection (Ch 6.1)
- **Sun 2/13**: **Homework 2 due**, before 11:59 PM to Canvas.

Week 06, 02/14 - 02/20:

- **Tue 2/15**: K-fold Cross Validation (Ch 5.1.3) / LASSO (Ch 6.2)
- **Thu 2/17**: Continued topics

Week 07, 02/21 - 02/27:

- **Tue 2/22**: **Quiz 2** Logistic regression, MLE (Ch 4.1-4.3.4)
- **Thu 2/24**: Continued topics
- **Fri 2/25**: **Homework 3 due**, before 11:59 PM to Canvas.

Week 08, 02/28 - 03/06:

- **Tue 3/1**: Spring break
- **Thu 3/3**: Spring break

Week 09, 03/07 - 03/13:

- **Tue 3/8**: Classification (ROC, AUC, FDR). Bayes rule
- **Thu 3/10**: Lasso

Week 10, 03/14 - 03/20:

- **Tue 3/15**: Text mining
- **Thu 3/17**: Leeway
- **Sun 3/20**: **Homework 4 due**, before 11:59 PM to Canvas.

Week 11, 03/21 - 03/27:

- **Tue 3/22**: Break
- **Thu 3/24**: Decision trees (Ch 8.1)

Week 12, 03/28 - 04/03:

- **Mon 3/28**: **Midterm Exam** 7:00 - 9:00 PM.
- **Tue 3/29**: Bagging/Random Forest (Ch 8.2)
- **Thu 3/31**: Random Forest (Ch 8.2)

Week 13, 04/04 - 04/10:

- **Tue 4/5**: Boosting (Ch 8.2)
- **Thu 4/7**: Leeway
- **Sun 4/10**: **Homework 5 due**, before 11:59 PM to Canvas. –>

Week 14, 04/11 - 04/17:

- **Tue 4/12**: Neural Network/Deep Learning
- **Thu 4/14**: Neural Network/Deep Learning

Week 15, 04/18 - 04/24:

- **Tue 4/19**: Neural Network/Deep Learning
- **Thu 4/21 (Last Class)**: **Quiz 3**. Last class: Final quiz

Week 16, 04/25 - 05/01:

- **Fri 4/29**: Data Science Live (DSL), 5:00 - 7:00 PM
- **Sun 5/1**: **Final project due before 11:59 PM to Canvas**