Machine Learning for Social Scientists
Political Science 150B/355B
Tuesday, Thursday 12–1:20pm
Building 60, Room 109
Ways: Applied Quantitative Reasoning

Instructor: Dr. Rochelle Terman, Postdoctoral Fellow, Center for International Security & Cooperation

Office: Encina Hall West, Room C225

Contact: rterman@stanford.edu.

Office Hours: Tuesdays, 3–5pm.

TA: Haemin Jee, Ph.D. Candidate, Political Science

Contact: hjee@stanford.edu

Office Hours: Tuesdays, 2–4pm.

TA: Tongtong Zhang, Ph.D. Candidate, Political Science

Contact: ttzhang7@stanford.edu

Office Hours: Mondays, 3–5pm.

Social scientists increasingly use large quantities of data to make decisions and test theories. For example, political campaigns use surveys, marketing data, and previous voting history to optimally target get out the vote drives. Governments use social media to track the extent of natural disasters. And political scientists use massive new data sets to measure the extent of partisan polarization in Congress, the sources and consequences of media bias, and the prevalence of discrimination in the workplace. Each of these examples, and many others, make use of statistical and algorithmic tools that distill large quantities of raw data into useful quantities of interest.

This course introduces techniques to collect, analyze, and utilize large collections of data for social science inferences. The ultimate goal of the course is to introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely. In achieving this ultimate goal, students will also:

1) Learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems.

2) Develop their programming abilities in `R`.

3) Be introduced to substantive problems and apply the techniques from the course.

4) Be able to learn independently and tackle more advanced topics and challenges in data analysis.

Prerequisites

Ideally students will have taken 150A or the equivalent. If you have any questions about preparing for the class, please talk to me.

Course Websites

We will be using **Canvas** to disseminate lecture notes, code, and data. Homework assignments will also be distributed and submitted through Canvas.

We will be using **Piazza** for communication. You may ask questions (anonymously if you wish) through Piazza and an instructor will respond promptly. You are also encouraged to answer one another's questions and engage in peer-to-peer learning. If you have a question related to course materials or concepts, please use Piazza before email.

Please sign up for the Piazza site here: `piazza.com/stanford/winter2018/polisci150b350b`

Evaluation

Students will be evaluated across five areas.

1. Homework : 35% of final grade. Students will be asked to complete five homework assignments. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for applied work. Portions of the homework completed in R should be submitted using R markdown, a markup language for producing well-formated HTML documents with embedded R code and outputs. R markdown requires installation of the `knitr` package. We recommend using Rstudio, a user interface for R, which is set up well for the creation of R markdown documents.

More about `RStudio` can be found here:
http://www.rstudio.com/

`R Markdown` can be found here:
http://rmarkdown.rstudio.com/

Students are encouraged to collaborate on problem sets together, but must write up their problem sets on their own.

We'll give the assignments on the following schedule.

HW 1 Assigned 1/11, due 1/23

HW 2 Assigned 1/23, due 1/30

HW 3 Assigned 1/30, due 2/8

HW 4 Assigned 2/15, due 2/22

HW 5 Assigned 2/22, due 3/6

2. Group Project : 15% of final grade. Each group will be assigned a topic from a list of popular machine learning tools. You'll work together to learn about the tool and present a broad overview to the class. Presentations will take place on 3/13.

3. Midterm Exam : 20% of final grade. Students will complete one in-class midterm exam. The exam will be held during class time on February 13.

4. Final Exam : 20% of final grade. Students will complete an in-class final exam. The exam will be held 12:15pm-3:15pm on March 22.

5. Participation : 10% of final grade. Students can earn participation through attending and asking questions in class, posting on piazza, and actively participating in weekly section.

Grading Policy

You can expect to receive a grade that accurately reflects the work you submit. There is no curve. Final grades are taken from many assignments, so no one task will wholly determine your grade.

Academic Dishonesty: I follow a zero-tolerance policy on all forms of academic dishonesty. All students are responsible for familiarizing themselves with, and following, Stanford's policies regarding proper student conduct. Being found guilty of academic dishonesty is a serious offense and may result in a failing grade for the assignment in question, and possibly for the entire course.

Late Policy: Submission deadlines for written assignments are inflexible. Extensions or incompletes will be given only to students with a **documented** emergency or illness. Please submit requests for rescheduling as far in advance of the scheduled date as possible.

Grade Disputes: All grades in this class are final. Because there are many assignments, there is much room for improvement if you receive a less-than-stellar grade on early assignments. I am always happy to meet and discuss general assignment and participation strategies. I will also meet to review past assignments, but only with the aim of helping you improve on the next.

Books

There is no required book for the course. We will post readings to coursework that will draw on other textbooks and popular writing that draws on machine learning approaches.

While there are no required texts, you might consider the following books as useful references.

ESL: Hastie, Trevor. Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* A classic and extensive treatment of machine learning concepts.

ISL: Gareth, Hastie, and Friedman. *An Introduction to Statistical Learning: With Applications in R.* Covers many of the same topics as ESL, but concentrates more on the applications of the methods and less on the mathematical details.. You can read the book for free here: `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf`

Class Outline

1/9 : 1. Introduction

Supervised Learning

1/11 : 2. Simple Regression

1/16 : 3. Multiple Regression

1/18 : 4. Classification

1/23 : 5. Logistic Regression

1/25 : 6. LASSO 1

1/30 : 7. LASSO 2 & Cross Validation

2/1 : 8. K Nearest Neighbor

2/6 : 9. Text Preprocessing

2/8 : 10. Distinctive Words

2/13 : **Midterm**

2/15 : 11. Dictionary methods

Unsupervised Learning

2/20 : 12. Text Geometry and Distances

2/22 : 13. Clustering

2/27 : 14. Topic Models

3/1 : 15. PCA

3/6 : 16. TBD

3/8 : *The Politics of Machine Learning* (Guest Speaker)

3/13 : Group Presentations

3/15 : Review