Introduction to Machine Learning
Political Science 150B/355B
Tuesday, Thursday 10:30am-11:50am
Ways: Applied Quantitative Reasoning

Instructor: Justin Grimmer, Political Science Department

Office: Encina Hall West, Room 414

Contact: jgrimmer@stanford.edu, 617-710-6803. Gchat; justin.grimmer@gmail.com

Office Hours: Schedule online at www.wejoinin.com/justingrimmer

TA: Bobbie MacDonald

Contact: bmacdon@stanford.edu

Office Hours: 1130-1

R TA: Tongtong Zhang

Contact: ttzhang7@stanford.edu

Office Hours: 1-3 pm, Friday, Encina West 427

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` and be introduced to `Python`.

3) Be introduced to substantive problems and participate in challenges applying the techniques from the course.

Prerequisites

Ideally students will have taken 150A or the equivalent. If you have any questions if you're prepared for the class, please talk to me.

Evaluation

Students will be evaluated across three areas.

Homework 30% of final grade. Students will be asked to complete four 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, an IDE 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 will also be introduced to `Python`. We will provide you with more introduction about that language as we introduce it.

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.

Homework 1 Assigned 1/12, due 1/20

Homework 2 Assigned 1/19, due 1/27

Homework 3 Assigned 2/9, due 2/17

Homework 4 Assigned 3/1, due 3/9

Contests 30% of final grade. We will complete two contests during the course. The contests will allow students to apply the techniques learned in the course to real problems. The teaching staff will introduce the challenges, rules, and evaluation of the submissions as we introduce the specific details of each challenge. The challenges will be the following:

Challenge 1 Predicting the Iowa Caucus Results (Very Challenging for many reasons!). Submissions will be due at 2/1 at 12 midnight (that is, the midnight before the Iowa Caucuse)

Challenge 2 Analyzing political texts with unsupervised models. Submissions will be due at the start of class on 3/1.

Midterm Exam 10% of final grade. Students will complete an in-class mid-term exam.

Final Exam 25% of final grade. Students will complete a cumulative in-class final exam.

Participation 5% of final grade. Students can earn participation through attending and asking questions in class, posting on piazza—our page is here
`https://piazza.com/stanford/winter2016/polisci150b355b/home`
and actively participating in weekly section.

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.

Recommended/Reference Books

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

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. A slightly more advanced text, but an excellent treatment of machine learning methods.

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.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. A more computer science oriented treatment of machine learning, with more extensive treatment of the estimation techniques used for machine learning methods.

Class Outline

Week 1, 1/5 Introduction and a machine learning focus on regression

Unit 1 Supervised Learning

Week 2, 1/12 : Classification and generalized linear models

Week 3, 1/19 : LASSO and Ridge regression

Week 4, 1/26 : Evaluating and selecting models

Week 5, 2/2 : Ensembles of Classifiers

Midterm: 2/4 : In class midterm exam.

Unit 2 Unsupervised Learning

Week 6, 2/9 : Clustering

Week 7, 2/16 : Topic Models for Text Analysis

Week 8, 2/23 : Factor Analysis, Principal Components, and Multidimensional scaling

Networks

Week 9, 3/1 : Network Models and Page Rank

Week 10, 3/8 : Industry guest speaker and review