POLISCI 251A: Introduction to Machine Learning for Social Scientists *

Instructor: Edgar Franco Vivanco

Summer, 2018

email: edgarf1@stanford.edu
www.edgarfrancovivanco.com
Location: Building 160, Room 124
Class Hours: Mon, Wed 1:30-2.50pm
Office hours: TBD
Class Units: 4

TA: Jesse Yoder (yoderj@stanford.edu)
Section Times: Th 10:30-11:20am
TA: Haemin Jee (hjee@stanford.edu)
Section Times: Th 4:30-5:20pm

Overview

Can we predict elections? How do we identify fake news? Can social media analysis anticipate protests? To answer questions like these, social scientists increasingly use large quantities of raw data combined with statistical and algorithmic tools. 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 these methods widely. Students will leave the course equipped with a broad understanding of machine learning and on how to continue building new skills.

This is an introductory course, so the lectures and problem sets will be focused on the intuition and the mechanics behind machine learning concepts rather than the mathematical fundamentals.

Prerequisites

There are no formal prerequisites for the course, but calculus and introductory statistics are **strongly** recommended. Students are not expected to have any programming knowledge, and the course will be centered around bite-size assignments that will help build R coding and statistical skills from scratch. The course will be centered around assignments that will introduce students to programing. If you have any questions about preparing for the class, please talk to me.

Course Objectives

In this course students will:

*This course is based on **POLISCI 150B** designed by Justin Grimmer, with some modifications implemented by Rochelle Terman.

1. Learn about the core concepts in machine learning and statistics, as well as their applications to solve pressing social problems such as electoral fraud, discrimination and the proliferation of fake news.
2. Develop their programming abilities in the R language.
3. Familiarize themselves with the applied literature in the topic.
4. Be able to learn independently and tackle more advanced topics and challenges in data analysis.

Learning approach

In this course I will use a semi-flipped classroom approach in which first I will introduce a concept, then each student will work individually in some code related to the concept. Then, we will go over the code together as a group. While working on the code, students are expected and encouraged to ask questions to the instructors and to other students. Please, bring your own laptop and download R and R studio in advance.

R can be found here: https://www.r-project.org/

R studio can be found here: https://www.rstudio.com/products/rstudio/download/

Grading Policy

Students will be evaluated using the following proportions:

- **15%** of your grade will be determined by a class midterm exam. The exam will be held during class time on July 23th.
- **40%** of your grade will be determined by 5 problem sets (8% each). The assignments are intended to expand upon the lecture material. 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.

  Students are encouraged to collaborate but will be responsible for writing their own R code, producing their own statistical output, and writing their own interpretations of what the resulting estimates mean from a substantive standpoint.

  Problem sets will be submitted via *Canvas*, and will be graded for three criteria: correct discussion of concepts; correct output of statistical code; and code "style," meaning how well commented and explained the submitted code is. Problem sets will not be accepted after the time at which they are due.
- **15%** Final project. 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. Each group will present their project during the last session.
- **20%** Final exam. The final exam will be cumulative and will test knowledge developed throughout the course.
- **10%** Students can earn participation through attending and asking questions in class and the labs. During some lectures we will discuss briefly a paper on applications of Machine Learning to Social Sciences. Students can also earn participation making meaningful contributions to these discussions.

Course Websites

We will be using Canvas to disseminate lecture notes, code, and data. Homework assignments will also be distributed and submitted through Canvas. Materials also will be available via `github`.

Labs

A Teaching Assistant (TA) will hold weekly labs to expand on lecture material and clarify particular questions. More information about lab schedules to come.

Schedule and weekly learning goals

The schedule and assigned readings are tentative and subject to change. The learning goals below should be viewed as the key concepts you should grasp after each week.

Supervised Learning

Week 1, Jun 25-29

Topics:

- Introduction. What is Machine Learning? A conceptual approach
- Intro to R *Homework 1 posted*.

Week 2, Jul 2-6

- Describing and relating variables *Homework 1 due*.
- Intro to Regression Analysis *Homework 2 posted*.

NOTE: No class on July 4th.

Week 3, Jul 9-13

- Machine Learning via OLS.
- Machine learning via GLS (Classification) *Homework 2 due, Homework 3 posted*.

Week 4, Jul 16-20

Topics:

- Comparing classification Methods
- Error checking and resampling *Homework 3 due*.

Week 5, Jul 23-27

Topics:

- **Midterm**
- Midterm Review and Intro to regularization

Week 6, Jul 30-Ag 3

- LASSO
- Decision Trees *Homework 4 posted*.

Unsupervised Learning

Week 7, Aug 6-10

Topic:

- Intro to Unsupervised Learning, Principal Components *Homework 4 due*.
- Clustering *Final Project Assigned and Homework 5 posted*.

Week 8, Aug 13-16

- Final review session and guest speaker. *Homework 5 due*.
- Group Presentations.
- Final exam (Date TBA)

Books

The structure of the course will follow **ISL.** Gareth, Hastie, and Friedman. *An Introduction to Statistical Learning: With Applications in R*. This book 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

In addition, I will post additional readings that will draw on other textbooks and popular writing.

Advanced students and those interested on the mathematical details of each method are advised to consult:

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

Course Policies

During Class

I understand that the electronic recording of notes could be important for class and so computers will be allowed in class. However, I strongly recommend to keep your computers closed during the non-coding portion of the class, and to take notes by hand. In general, students should refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course.

Attendance Policy

This course has a relatively fast pace, so missing even one session would be detrimental for understanding the material. Therefore, attendance is expected in all lecture sections. Note that attendances is also part of the grade.

Policies on Late Assignments

Late assignments are not accepted under any circumstance. Since there are many opportunities to obtain points during the quarter, I strongly recommend to get as much as you can in each assignment and handle it -even if incomplete.

Grade Policy

You can expect to receive a grade that accurately reflects the work you submit. There is no curve and all grades in this class are final. Please, do not request any grade revision for your homework grades. I am always open to discuss your homework but only to help you improve in the next assignment.

Academic Integrity and Honesty

Students are expected to adhere to the Stanford Honor Code (http://studentaffairs.stanford.edu/communitystandards/policy/honor-code) at all times. Collaboration is encouraged on problem sets, but students must identify the students with whom they collaborated at the top of their submitted problem set.

Accommodations for Disabilities

Students who may need an academic accommodation based on the impact of a documented disability must initiate the request with the Office of Accessible Education (OAE) as son as possible. More information here: http://studentaffairs.stanford.edu/oa.