Model Based Inference and Machine Learning
Political Science 450c, Spring 2018
Mon and Wed 9:30 AM - 11:20 AM GSL
Section: Friday 9:30am-11:20am

Instructor: Jens Hainmueller

Office: Immigration Policy Lab, Encina West Basement

Contact: jhain@stanford.edu

Office Hours: Please schedule an appointment at `http://www.meetme.so/jhain`

TA: Mathilde Emeriau

Contact: memeriau@stanford.edu

Office Hours: Mondays, 4:30pm - 6:30pm at Encina 467

TA: Matt Tyler

Contact: mdtyler@stanford.edu

Office Hours: Tuesdays, 9:00am - 11:00am at Encina 467

POL 450C continues the graduate methods sequence. In this quarter you will learn about model based theories of inference in political science building up to their application for Machine Learning tasks. Topics covered will include likelihood-based inference, generalized linear models, discrete choice models, regularization, missing data, latent-variable models, and (a brief introduction to) optimization approaches.

One primary goal of this course is to teach students about a wide array of models that are widely used across political science, but closely related to methods that you've already seen in the methods sequence. A closely related goal will be to assess what the additional methods contribute beyond the methods that you have already learned. Students will learn how to both interpret the output of these more complicated models and to better assess claims about the benefits the models provide. A second primary goal of the course is to acquaint students with many of the most important recent advances in machine learning methods. Many of these methods build on the models we introduce in the first part of the course, while others require new intuition. We will see that machine learning methods exploit the increases in computer power to make better use of the available data. Throughout the course, we will be careful to clarify our inferential goals and to inquire when a model based approach is likely to improve (or harm!) causal inferences, descriptive inference, or facilitate exploration.

Our secondary goal will be to continue developing your programming and mathematical proficiency. Students will be pushed to write code that accomplishes more complicated tasks using code that is cleaner and more efficient. And some models will be presented at a slightly higher level of mathematical abstraction—both to convey the content of the models and to prepare students to be sophisticated users of the best current statistical models.

Prerequisites

Students are required to have taken math camp, POL 350A, and POL 350B. Special permission from the instructor is required if you have not taken the prerequisites.

Course Website

The course website is located on Piazza at:
piazza.com/stanford/spring2018/450c/home

You can sign up on the Piazza course page directly from the above address. We will distribute course materials—including readings, lecture slides and problem sets—on this website. There is also a question-and-answer platform that is easy to use and designed to get you answers to questions quickly. It supports LaTeX, code formatting, embedding of images, and attaching of files. We encourage you to ask questions on the Piazza forum in addition to attending recitation sections and office hours.

Using Piazza will allow students to see and learn from other students' questions. Both the TAs and instructors will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. A student's respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructors or TAs* (unless they are of personal nature) — we will not answer them!

Evaluation

Students will be evaluated across five areas.

Homework 25% Students will be asked to complete a weekly homework assignment. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for their work.

Midterm Exam : 20% Students will be asked to complete a closed book, pencil and paper only midterm exam. It will take place in class on May 16.

Final Exam : 25% Students will be asked to complete a closed book, computer based final exam. It will take place during the final exam week. You are required to work on this exam alone.

Replication project 25% Working in pairs students will be asked to complete a replication and reanalysis of a published political science article. The article should use a statistical technique from either 350A, 350B, and 350C along with quantitative data. Students should consult with the instructor and TA about the choice of paper. Students will be graded on their ability to productively evaluate the original modeling choices in the paper and to provide useful extensions. Students will present their replication projects to the class.

Participation 5% Students are expected to attend each class and to ask questions regularly.

Class Dates

- First day of class is Wed April 4 (no class Mo April 2)
- No class Mon May 28 (Memorial Day).
- Midterm May 16
- Replication Presentation June 6

Books

The following are required books for the course

- Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models.* Wiley. (Hereafter AA)
- Wasserman, Larry. 2013. *All of Statistics: A Concise Course in Statistical Inference.* Springer. (Hereafter AS, available electronically from the library.)
- Wasserman, Larry. 2006. *All of Nonparametric Statistics* (Hereafter ANS, available electronically from the library.)
- Bertsekas, Dimitri P and Tsitsiklis, John. Introduction to Probability Theory (Hereafter BT. You purchased this book for math camp.)
- Hastie, Tibshirani, and Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd edition. (Available electronically from the authors.)

We will supplement the readings with other books as appropriate.

Class Outline (Topics and Schedule Subject to Change)

Probability Theory: A Refresher

- AS Chapter 1-5 (pg 3-82)
- BT Chapter 1-5
- Math camp slides

Likelihood Theory of Inference

- AS Chapter 9
- Degroot and Schervish 6.1-6.5 (Hand out)

Models for normally distributed outcomes

- AA Chapter 4 (Chapters 2-3 for background)

Logit and Probit Models for Binary Outcomes

- AA Chapter 5
- AS Chapter 13

Bootstrap, Monte Carlo, and Delta Method

- ANS Chapter 3
- AS Chapter 9.9
- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation.". *American Journal of Political Science.* 44,2. 347-361.

Ordered Probit/Logit

- Gelman, Andrew et al. 2008. "A Weakly Informative Default Prior for Logistic and Other Regression Models". *The Annals of Applied Statistics.* 2, 4. 1360-1383.
- AA Chapter 6.2
- `http://web.stanford.edu/class/polisci203/ordered.pdf`

Mulitnomial Logit/Multinomial Probit

- AA Chapter 6.1

Event count models

- AA Chapter 7

Hypothesis Testing in GLMs

- AS Chapter 10.1-10.3, 10.6
- AA Chapter 4.1-4.3

GLMs and model checking

- AA Chapter 4.4-4.6
- AS 10.8
- ESL. 7.1-7.7 (Handout)

Nonparametric Regression and Generalized Additive Models

- AS Chapter 6
- Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models". *American Journal of Political Science* 42, 2. 596-627.

Tree Based Learning Models

- Machine Learning, a Probabilistic Perspective Chapter 16 (Handout)

LASSO, Ridge and Regularization

- ESL, 3.4.2 (Handout)
- Introduction to Statistical Learning (James, Whitten, Hastie, Tibsharani) (available online) Chapter 6

Kernel Regularized Least Squares

- ESL, 3.4.1.
- Machine Learning, a Probabilistic Perspective Chapter 8 (Handout)
- Hainmueller, Jens and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach" *Political Analysis*. 22, 2. 143-168.

Missing Data and Multiple Imputation

- Little, Roderick JA, and Donald B. Rubin 2014. Statistical analysis with missing data.