Model Based Inference and Introduction to Machine Learning
Political Science 450c, Spring 2019
Monday, Wednesday 900-1050am

Instructor: Justin Grimmer, Political Science Department

Office: Encina Hall West, Room 416

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

Office Hours: My door is almost always open during normal business hours. If you absolutely need to see me at a particular time, please schedule an appointment.

TA: Matt Tyler, Political Science Department

Contact:

Office Hours:

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 for discrete choice, event count models, regularization, 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 massive 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. A third primary goal will be to acquiant students with modern research in political science.

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 the 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 450A, and POL 450B. Special permission from the instructor is required if you have not taken the prerequisites.

Evaluation

Students will be evaluated across five areas.

Homework 30% 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. 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 makes it easy to create R markdown documents.
More about `RStudio` can be found here:
http://www.rstudio.com/

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

Midterm Exam : 15% Students will be asked to complete a closed book, pencil and paper only midterm exam. It will take place during class.

Final Exam : 25% Students will be asked to complete a closed book, computer based final exam. It will take place during the final exam time for the course.

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 450A, 450B, and 450C 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.

Participation 5% Students are expected to attend each class and to ask questions regularly. To encourage questions and discussion we will use slack.

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

4/2, Probability Theory: A Refresher

- AS Chapter 1-5 (pg 3-82)

- BT Chapter 1-5

- Math camp slides

4/4, Likelihood Theory of Inference

- AS Chapter 9

- Degroot and Schervish 6.1-6.5 (Hand out)

4/9, Models for normally distributed outcomes

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

4/11, Logit and Probit Models for Binary Outcomes

- AA Chapter 5

- AS Chapter 13

4/16, 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.

4/18, Near-Perfect Separation and Ordered Probit

- Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models". *Political Analysis*. 13, 2. 157-170.
- 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`

4/23, Mulitnomial Logit/Multinomial Probit

- AA Chapter 6.1

4/25, Event count models

- AA Chapter 7

4/30, Duration Models

- Klein, John and Moeschberger, Melvin L. *Survival Analysis: Techniques for Censored and Truncated Data*. Chapter 2, 12 (Hand out)

5/2, Hypothesis Testing in GLMs

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

5/7, GLMs and model checking

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

5/9, AIC/BIC and Nonparametric Measures of Model fit: Cross validation

- ESL 7.10 (Handout)

Midterm: 5/14

5/16, Principal Component Analysis (Kernel PCA in Slides Appendix)

- Machine Learning, a Probabilistic Perspective 12.2 (Handout)
- Introduction to Statistical Learning (James, Whitten, Hastie, Tibsharani) (available online) 10-10.2

5/21, LASSO, Ridge and Regularization

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

5/23, No Class, Work on Projects

5/28 SVMs, Kernel Trick, and KRLS

- 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.

5/30 Boosting/Bagging/Wisdom of the Crowds

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

6/4, Nonparametric Density Estimation and regression

- AS Chapter 20
- ANS Chapter 6
- 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.