Political Science 597D: Robust Models, Exploratory Models and Machine Learning

Tuesday 1-4pm Pond 236

Spring 2011

James Honaker
Pond 317

Phone: (814) 865 6901
tercer@psu.edu

1 Goals of This Course

Quantitative social science is a winnowing process. Important results are subjected to constant reassessment. Some analyses are shown to be fragile; predictions are heavily dependent on seemingly minor model choices. Some analyses are rejected entirely in favor of alternate variable choices or more sophisticated specifications. Some analyses are inordinately robust to all of these challenges and broadly replicable across model and variable choices, or even generalize to other data universes. Determining the robustness of your statistical model, letting the data guide you to a robust specification, and knowing how to make predictions when multiple models cannot be rejected, are important foundations to quantitative results that will withstand repeated replication and extension. This course focuses on automated techniques for these goals, broadly categorized as *machine learning* or *statistical learning* or even *data mining.*

2 Background and Foundation

Much of this class is really a tangent to the trajectory of the traditional visualization of the methods sequence. The methods sequence is often described as building increasingly sophisticated and elaborate mousetraps tailored to the perculiarities of individual breeds of mice. However, throughout the statistics coursework you take, as well as related research design classes and more focused applied classes, you also receive training in the pragmatics of sound quantitative investigation. The meta questions of how to code variables, how to decide to include and exclude variables, and how to deal with conflicting literatures, papers or results, are all necessary skills we develop. This class focuses on developing these questions.

About half the members of this class come in having taken a class in maximum likelihood methods, and half will be concurrently taking an generalizedd linear models class. This should not fundamentally prove a problem. The baseline model throughout this class is ordinary linear regression, with a few weeks related to logistic regression in the middle of the class. Those who know more complicated statistical models should understand that most instances of “regression” that appear in this class are simply placeholders for “whatever model is most appropriate for this dependent variable and data generating process.” For parsimony, and so as to more easily highlight what is novel in each technique, we will stick with assuming linear regression is a reasonable first pass model for the variables at hand. Those who have already taken MLE will be familiar with some of the topics we cover, if previously set out from a different vantage point and promoted from a different set of virtues.

3 Text

A text we will be using is the second edition of the very influential *Elements of Statistical Learning* by Hastie, Tibshirani and Friedman (MLE students may recognize the first two authors as also writing a very influential work on generalized additive models, with which there is some overlap). This text is not in the bookstore. If you want the standard edition (hardcover, color figures), it is discounted by a third on Amazon, but still expensive ($90 down to $60). The Penn State library system has electronic rights to this Springer series, so you can read online, or download in PDF the entire text for free. If you want a paper copy, but don't want to print out the PDF, you can buy a black-and-white, paperback version for $25 including shipping.

`http://www.springerlink.com/content/978-0-387-84857-0`

There will be a small number of readings, mostly applications of some of these methods in political science. These are intended primarily as catalysts for thinking about how these methods could be used in the subfields you are interested in, while also partly to reassure you that these methods are used in political science.

4 Assignments and Final Paper

There will be a short assignment almost every week. Write-ups can generally be as sparse as a couple pages. They are almost "quantitative response papers." Lecture will be roughly one hour of new material, one hour of pragmatically applying the idea to some political science problem, and a final hour reviewing and discussing the assignments that came in the previous week. For this reason, assignments have to come in on time, while at the same time, no individual assignment will be a large project. Assignments are due absolutely no later than Monday midnight the day before class. There will be an assignment every week; you can miss two.

The assignments will use *R*, not because the models are fundamentally hard, but because Stata is not flexible enough to build or iterate the simple models that we need. Those familiar with *R* will naturally have a small advantage in the ease with which they can begin an assignment and the tangents they can explore, but generally I will give you a foundation of code from which to begin, and the exercise should be mostly conceptual exploration rather than statistical programming.

Finally, the assignments should be TeX'd up. It's unfortunate, but a norm in the profession, that TeX ability is a weak signal of technical competence among the methods community. If you are working with an uncommon or challenging technical method, you want your audience or reader to start with a prior belief that you are up to the task. I'll provide some basic LaTeX templates from which you can insert text and figures for the first assignments.

Turn in an electronic copy of your assignments to `tercer@psu.edu`. I will be in town Monday afternoon through Thursday evening, but because I am less available, I will attempt to be consistently available by email.

There is no major final paper for this class, rather there will be a larger than usual final homework assignment over the last few weeks of term. In this, you will take some assignment, or combination of assignments, and apply it in more detail to a dataset of your own choosing. This should be a capstone to the set of assignments you have turned in over the term, and perhaps serve as a launching point for a future paper or project, or be part of a paper for another class

5 Organization of Topics

The structure of this class can be set out from different organizational principles. Thematically, there are four major theoretical topics. These are set out below:

- Variable Selection
  1 Subset, Stepwise, and Stagewise Regression. Information Criteria.
  3 Shrinkage estimators: Priors, Ridges, Lasso and Principal Components.
- Inference from Resampling
  2 Training/Validation/Test partitions and Out-of-sample Forecasting
  4 Bootstrapping
  9 Bagging
- Classification
  5 Generalized Additive Models, Kernels, Splines
  6 Nearest-Neighbor Matching
  7 Naive Bayes
  11 Trees
  13 Support Vector Machines
  14 Neural Nets
- Model Averaging
  8 Bayesian Model Averaging
  10 Boosting
  12 Random Forests

The numbers correspond to roughly the order in which we will move through these topics. So it's worth noticing that we'll oscillate between these different major theorectical themes. If we set out these topics roughly chronologically, we could broadly categorize them with more pragmatic, applied research oriented labels:

- Robust Model Estimation

  1 Subset, Stepwise, and Stagewise Regression. Information Criteria.
  2 Training/Validation/Test partitions and Out-of-sample Forecasting
  3 Shrinkage estimators: Priors, Ridges, Lasso and Principal Components.
  4 Bootstrapping
  5 Generalized Additive Models, Kernels, Splines

- Semi-Parametric Classification

  6 Nearest-Neighbor Matching
  7 Naive Bayes
  8 Bayesian Model Averaging
  11 Trees

- Robust Forecasting

  9 Bagging
  10 Boosting
  12 Random Forests
  13 Support Vector Machines
  14 Neural Nets

A third way we could divide the material is to call weeks 1 through 8, *things occasionally used in political science* and weeks 9 through 15 *things rarely used in political science.* Which of these frameworks you prefer to use to mentally divide the topics depends on how good a person you are.

Academic Dishonesty

The Department of Political Science, along with the College of the Liberal Arts and the University, takes violations of academic dishonesty seriously. Observing basic honesty in one's work, words, ideas, and actions is a principle to which all members of the community are required to subscribe.

All course work by students is to be done on an individual basis unless an instructor clearly states that an alternative is acceptable. Any reference materials used in the preparation of any assignment must be explicitly cited. Students uncertain about proper citation are responsible for checking with their instructor.

In an examination setting, unless the instructor gives explicit prior instructions to the contrary, whether the examination is in class or take home, violations of academic integrity shall consist but are not limited to any attempt to receive assistance from written or printed aids, or from any person or papers or electronic devices, or of any attempt to give assistance, whether the one so doing has completed his or her own work or not. Lying to the instructor or purposely misleading any Penn State administrator shall also constitute a violation of academic integrity.

In cases of any violation of academic integrity it is the policy of the Department of Political Science to follow procedures established by the College of the Liberal Arts.

Disabilities

The Pennsylvania State University encourages qualified people with disabilities to participate in its programs and activities and is committed to the policy that all people shall have equal access to programs, facilities, and admissions without regard to personal characteristics not related to ability, performance, or qualifications as determined by University policy or by state or federal authorities. If you anticipate needing any type of accommodation in this course or have questions about physical access, please tell the instructor as soon as possible. Reasonable accommodations will be made for all students with disabilities, but it is the student's responsibility to inform the instructor early in the term. Do not wait until just before an exam to decide you want to inform the instructor of a learning disability; any accommodations for disabilities must be arranged well in advance.