Summer 2018 Workshop in Political Methodology: Machine Learning

University of Colorado Boulder

Summer 2018

Lecture Time:	May 14-17, 9:30-12:00
Lab Time:	1:00-2:30
Location:	KTCH 1B31

Instructor:	Dr. Andrew Q. Philips
Office:	KTCH 144
Email:	andrew.philips@colorado.edu

COURSE DESCRIPTION: Machine learning is a growing field in the social sciences. Although some of the "big data revolution" is hype, there are a number of interesting and worthwhile tools that political scientists can add to their toolkit for data exploration, visualization, prediction, classification, and modeling.

This four-day course is designed to introduce students to the world of machine learning. We will start with simple supervised learning techniques—such as OLS—as well as those that may be less familiar, like LASSO and ridge regression. We then move to tree based approaches that are able to handle a variety of non-linear relationships. We may also cover some more novel topics, such as unsupervised learning techniques used to assess model fit, and resampling. We will place strong emphasis on being able to implement and interpret the output of these models through the use of labs.

By the end of the course, students should be comfortable with a variety of supervised and unsupervised learning techniques, and should be able to apply them to their own research questions.

PREREQUISITES: At least Data I (and preferably Data II), i.e., an introductory regression course.

SOFTWARE: We will use R for this course. Although familiarity with R is not necessary, it is a plus. For those unfamiliar with this program, there are copious amounts of information available for free online. Please download both R (https://cran.r-project.org/) and RStudio (https://www.rstudio.com/) before the first class session.

RECOMMENDED TEXTS: Either text is suggested for the course. Hastie, Tibshirani, and Friedman is a more advanced text, but covers more topics. James et al. is more accessible and contains R code. Any additional readings will be made available to you on the first day of class or as needed.

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2011. (HTF 2011) "The elements of statistical learning: Data mining, inference, and prediction." Springer Series in Statistics. 2nd edition. ISBN: 978-0387848570.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. (JWHT 2013) "An introduction to statistical learning: With applications in R." Springer Series in Statistics. 1st edition. ISBN: 978-1461471370

TENTATIVE SCHEDULE:

Day 1: "Envisioning OLS as a loss function"

Introduction, supervised prediction and penalized regression

Suggested Readings:

- JWHT Chapter 2, 3 and 6.
- HTF Chapter 2, and 3.
- Grimmer, Justin. 2015. "We are all social scientists now: How big data, machine learning, and causal inference work together" PS: 80-83.
- Nickerson, David W. and Todd Rogers. 2014. "Political campaigns and big data" *The Journal of Economic Perspectives* 28(2): 51-73.
- Hindman, Matthew. 2015. "Building better models: Prediction, replication, and machine learning in the social sciences" *The Annals of the American Academy of Political and Social Science* 659(1):48-62.
- Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net" *Journal of the Royal Statistical Society* 67(2): 301-320.

Day 2: "How can we assess how well our model fits the data?"

Classification, assessing model fit, resampling

Suggested Readings:

- HTF Chapter 4 and 7.
- JWHT Chapter 4 and 5. and local kernel methods"*The Annals of Applied Statistics* 2(3):777-807.

Day 3: "Tree-based regression and classification approaches"

Classification and regression trees, bagging, boosting, and random forests

Suggested Readings:

- HTF Chapter 9, 10, 15 and 16
- JWHT Chapter 8
- Muchlinski, David, Davis Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing Random Forest with logistic regression for predicting class-imbalanced civil war onset data" *Political Analysis* 24(1): 87-103.
- Suzuki, Akisato. 2015. "Is more better or worse? New empirics on nuclear proliferation and interstate conflict by Random Forests" *Research & Politics*.
- Siroky, David S. 2009. "Navigating Random Forests and related advances in algorithmic modeling" *Statistics Surveys* 3:147-163.
- Varian, Hal R. 2014. "Big data: New tricks for econometrics" *Journal of Economic Perspectives* 28(2): 3-28.

Day 4: "Support vector machines, neural networks, and the frontiers of 'big data' today"

Unsupervised learning and various other topics

Suggested Readings:

- HTF Chapter 11, 12 and 14
- JWHT Chapter 9 and 10
- Borisyuk, Roman, Galina Borisyuk, Colin Rallings and Michael Thrasher. 2005. "Forecasting the 2005 general election: A neural network approach" *The British Journal of Politics and International Relations* 7(2)"199-209.
- D'Orazio, Vito, Steven T. Landis, Glenn Palmer and Philip Schrodt. 2014. "Separating the wheat from the chaff: Applications of automated document classification using support vector machines" *Political Analysis* 22(2):224-242.

STATEMENT ABOUT STUDENTS WITH DISABILITIES

The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services—either online at http://www.colorado.edu/disabilityservices/—or at the Center for Community, N200, 107 UCB.

To best accommodate students who may require alternative services, it is crucial that you contact me *early in the semester* if you need such accommodations.

HONOR CODE, COPYRIGHT, AND PLAGARISM STATEMENTS

"*On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance*"

The CU Honor Code is intended to uphold the intellectual reputation of the university by establishing trust among individuals regarding intellectual honesty. As the website states, "The Honor Code secures an environment where academic integrity can flourish and aims to install the principles of honesty, trust, fairness, respect, and responsibility as essential features of the University of Colorado Boulder campus". Violations of intellectual honesty include plagiarism, cheating, and the unauthorized use of materials, all of which erode trust among individuals. If you have any questions about this, please see me, the Honor Code website (http://www.colorado.edu/honorcode/), or the Honor Code Office (1B70 Regent Admin Building).

The handouts and lectures used in this course are copyrighted. By "handouts," I mean all materials generated for this class, which include but are not limited to syllabi, exams, in-class materials, and review sheets. Because these are copyrighted, you do not have the right to copy them or distribute them to others outside class, unless I expressly grant permission. In addition, I do not grant permission to tape class lectures.

Last updated: May 13, 2018