Machine Learning for Public Policy

Professor: Guillaume A. Pouliot
TAs: TBD
Problem Session: TBD
Office Hours: TBD

Course Description

This course is the third installment of the three-quarter core sequence of the Data Science Certificate at the Harris School of Public Policy. Students at Harris and in the College may enroll, with permission of the instructor, without having taken previous courses in the sequence. However, it is necessary for MPP students to take the full sequence in order to meet the necessary requirements of the Data Science Certificate.

The objective of the Data Science sequence is to train students to be successful and autonomous applied economists and data scientists in government and industry. In the first two courses of the sequence, students learned programming, as well as how to handle, summarize, and visualize modern datasets.

The objective of this course is to train students to be insightful users of modern machine learning methods. The class covers regularization methods for regression and classification, as well as large-scale approaches to inference and testing. In order to have greater flexibility when analyzing datasets, both frequentist and Bayesian methods are investigated.

Typical applications of the methods presented in this course include, but are not limited to: predicting restaurants' sanitation inspection scores, uncovering the determinants of recidivism, testing for judges' impartiality, and carrying out regression analysis and model selection using surveys with very many variables, such as the Current Population Survey.

Lectures

Section 1: Linear Regression

1. OLS: Revision and Large Scale-OLS
2. Regression Trees

Section 2: Bootstrap

3. The Bootstrap and Alternatives: Testing and Inference

Section 3: A/B testing

4. Design of Experiments
5. Permutation Tests
6. Permutation Tests with Compromised Randomization

Section 4: Bayesian and Monte Carlo Methods

7. MCMC Basics: Metropolis-Hastings as a Unifying Concept
8. MCMC Advanced Methods
9. Topic Models: Natural Language Processing

Section 5: Regularized Methods and Model Selection

10. Regularized Regression: Lasso, Ridge, and Friends
11. Regularized Regression: Generalized Linear Models
12. Model Selection: Cross-Validation, Information Criteria, MSE Estimation and Advanced Methods
13. Classification: Support Vector Machines and Friends

Section 6: Testing and Inference with Big Data

14: Inference with Big Data: Inference Post Model-Selection and Bootstrap Inference
15: Large-Scale Hypothesis Testing: False Discovery Rates

Section 7: Random Forests and Boosting

16. Random Forests
17. Boosting

Section 8: Deep Learning

18. Deep Learning Basics
19. Deep Learning Topics

Material

The main texts for this course are:

Efron, Bradley, and Trevor Hastie. *Computer Age Statistical Inference*. Vol. 5. Cambridge University Press, 2016.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Additional reading material pertaining to individual sections include:

Section 1

Weisberg, Sanford. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

Varian, Hal R. "Big data: New tricks for econometrics." *The Journal of Economic Perspectives* 28, no. 2 (2014): 3-27.

Athey, Susan, and Guido W. Imbens. "Machine learning methods for estimating heterogeneous causal effects." *stat* 1050 (2015): 5.

Section 2

Efron, Bradley, and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Section 3

Mead, Roger. *The design of experiments: statistical principles for practical applications*. Cambridge university press, 1990.

Kennedy, Fetter E. "Randomization tests in econometrics." *Journal of Business & Economic Statistics* 13, no. 1 (1995): 85-94.

Heckman, James J., Rodrigo Pinto, Azeem M. Shaikh, and Adam Yavitz. *Inference with imperfect randomization: The case of the Perry Preschool Program*. No. w16935. National Bureau of Economic Research, 2011.

Section 4

Hoff, Peter D. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. "An introduction to MCMC for machine learning." *Machine learning* 50, no. 1-2 (2003): 5-43.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Mcauliffe, Jon D., and David M. Blei. "Supervised topic models." In *Advances in neural information processing systems*, pp. 121-128. 2008.

Blei, David M., and John D. Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*, pp. 113-120. ACM, 2006.

Section 5

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, no. 3 (2011): 273-282.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*. CRC press, 2015.

Bühlmann, Peter, and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software* 33, no. 1 (2010): 1.

Claeskens, Gerda, and Nils Lid Hjort. *Model selection and model averaging*. Vol. 330. Cambridge: Cambridge University Press, 2008.

Efron, Bradley. "Estimating the error rate of a prediction rule: improvement on cross-validation." *Journal of the American Statistical Association* 78, no. 382 (1983): 316-331.

Efron, Bradley, and Robert Tibshirani. "Improvements on cross-validation: the 632+ bootstrap method." *Journal of the American Statistical Association* 92, no. 438 (1997): 548-560.

Smola, Alex J., and Bernhard Schölkopf. *Learning with kernels*. GMD-Forschungszentrum Informationstechnik, 1998.

Section 6

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*. CRC press, 2015.

List, John A., Azeem M. Shaikh, and Yang Xu. *Multiple hypothesis testing in experimental economics*. No. w21875. National Bureau of Economic Research, 2016.

Section 7

Schapire, Robert E., and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

Section 8

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.