Machine Learning PPHA 30545

Winter 2019

Instructor: Guillaume Pouliot

Group 1: Monday-Wednesday, 9:30-10:50

Group 2: Monday-Wednesday, 11:00-12:20

TAs: TBD

Instructor OH: TBD

TA OH: By appointment

Section: TBD

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.

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 first few lectures of the course are dedicated to a thorough review of matrix algebra, as it underpins much of the machine learning methods and theory in the course.

Course Policies:

No Laptop Policy: Laptop computers are not allowed in class.

Collaboration on Problem Sets: You are encouraged to collaborate on problem sets, but you should write your own code and your own solutions.

Distribution of Material: The slides will be distributed, but you should not let that deter you from doing the reading assignments. The material is covered in greater detail in the readings. The assigned reading cover the material in greater depth and should be considered as the reference.

Texbooks:

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

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Vol. 112. New York: springer, 2013.

Copyright:

Some lectures use course material developed by Matt Taddy, Professor of econometrics at Chicago Booth and VP chief economist for marketplaces at Amazon. The slides and problem sets designated as such are his property.

Grading:

Problem Sets: 50%

Midterm: 20%

Final: 30%

Course Outline:

Week 1: Matrix algebra review. Applications to multivariate linear regression and related topics in causal inference.

readings: Weak and many IV handout. Matrix algebra handout.

Week 2: Matrix algebra topics. Applications to principal component analysis and clustering.

Readings: Matrix algebra handout.

Week 3: Topics in regression: quantile regression.

readings: Koenker (2005) *Quantile Regression*. Chapters 1 and 2.

Week 4: Markov chain Monte Carlo methods.

readings: TBA.

Week 5: Data: Computing, plotting, FDR. Regression: A grand overview, linear and logistic.

readings: Intro and FDR: CASI p.1-11 and p.271-282

Logistic Regression: ISL p.127-138

Week 6: Model Selection: penalties, information criteria, cross-validation.

Treatment Effects: HD controls, propensity scores, bootstrap

readings: Bias-Variance Tradeoff : ISL p. 29-42

Out-of-Sample fit and Cross-Validation: ISL p.175-186 and/or CASI p. 208-232

Lasso: ISL p.203-227, note that p. 210-213 cover AIC, and/or CASI p.298-308

BIC: CASI p. 243-250

ROC: ISL p.147. For more background, consult

https://www.dataschool.io/roc-curves-and-auc-explained/

http://corysimon.github.io/articles/what-is-an-roc-curve/

the first pages of this optional reading may be helpful

http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf

Bootstrap: CASI p.155-177 and ISL p.187-190

Week 7: Classification: Multinomials, KNN, sensitivity/specificity

readings: Please read chapter 4 of ISL about Classification.

Week 8: Networks: co-occurance, directed graphs, Page Rank/ Clustering: Mixture models, k-means.. Factors: latent variables, PCA, PCR, and PLS.

readings: Please do the readings for factor models:

ISL: p. 373-383

Week 9: Trees: CART and random forests, ensembles

readings: TBA

Week 10: Text Mining: topic models, sentiment prediction, deep learning

readings: Latent Dirichlet Allocation (2003) Bei et al., Journal of Machine Learning Research.

The **Midterm** exam is on a date TBD

Academic Honesty:

1. Academic dishonesty will not be tolerated. Plagiarism on a problem set, at first offense, will result in a grade of zero on that problem set. If you commit a second offense, or plagiarise on an exam, you may receive an F in the class.
2. All the work must be your own.
    1. You are encouraged to discuss the problem set questions with classmates.
    2. However, you should write out your own solutions.