Machine Learning, Winter 2021, PPHA 30545

Professor: Guillaume Pouliot

Email : guillaumepouliot@uchicago.edu

Group 1: Tuesday-Thursday, 2:40-4:00

Group 2: Tuesday-Thursday, 4:20-5:40

Head TA: Rahim Rasool, rahimrasool@uchicago.edu

TAs: Maca Guzman Valenzuela, macaguzman@uchicago.edu

Nathan Dignazio, ndignazio@uchicago.edu

Qiwei Lin, qiweilin@uchicago.edu

Instructor OH: TBA

TA OH: By appointment

Lab section 1: Friday, 9:10-10:30am

Lab section 2: Friday, 10:50-12:10

The course is online.

This course is a high-level introduction to a selection of fundamental and modern machine learning methods. Each week presents, explores and applies a different family of methods. A wide array of methods is covered, and the objective of the course is to train students to carry out basic statistical machine learning analysis using these, and become informed and critical consumers of machine learning research.

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.

Course Policies:

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 readings cover the material in greater depth and should be considered as the reference.

Textbook:

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

Grading:

Problem Sets: 55%

Midterm quiz: 10%

Final quiz: 15%

Participation: 20%

Recall that Harris has a standard grade distribution of 1/8 of A's, 1/4 of A-'s, 1/4 of B's, and 1/8 of B-'s and below.

Problem sets:

Each problem set will have two sections. The first section presents problems to help students concretize their understanding of previously taught material, and the second section asks high-level questions about upcoming material to motivate the lecture and give students time to ponder upon the methodological questions which will be answered in the lecture. For full credit, the latter questions need to be answered thoughtfully, but not necessarily correctly.

Midterm quiz:

The midterm quiz will review theoretical material from the first-half of the class.

Final quiz:

The final quiz covers material from the whole course. A set of practice questions will be handed to the students a week before the exam.

Participation:

Participation is based on students reading ahead and coming to class with at least one prepared question at every lecture.

Outline:

Week 1: Introduction and key concepts of statistical machine learning

Tuesday, January 12: Introduction. Course overview. The decision theoretic basis of machine learning.
Readings: ISL, chapter 1

Thursday, January 14: Review of basic material. Simple regression. Basic Inference. Hypothesis testing. Basic multiple hypothesis testing.
Readings: ISL, chapter 2

Problem set 1 released, Friday, January 15

Weeks 2: Multivariate linear regression

Laboratory 1 released, Monday, January 18

Tuesday, January 19: Factorial models. Nonlinear transformations and interactions. Matrix notation. Geometry of least-squares.
Readings: ISL, chapter 3

Thursday, January 21: The regression function. Model misspecification. Best linear predictor. Identification and forecasting with singular design matrices.

Readings: ISL, chapter 3

Problem set 1 due, Monday, January 25

Weeks 3 and 4: Model Selection: Penalty function and resampling methods

Tuesday, January 26: Multiple hypothesis testing methods. Model Selection as multiple hypothesis testing. Uniformly valid inference.

Readings: handout 1

Thursday, January 28: False discovery rates and the Benjamini-Hochberg theorem. ROC curves.

Readings: handout 2

Problem set 2 released, Friday, January 29

Laboratory 1 due, Monday, February 1

Laboratory 2 released, Monday, February 1

Tuesday, February 2: The bootstrap, cross-validation and permutation tests.

Readings: handout 3

Thursday, February 4: Advanced topics. Improvements on the bootstrap. Limits of the bootstrap.

Readings: handout 3

N.B. End of material for the midterm.

Weeks 5: Priors, shrinkage, and regularization

Problem set 2 due, Monday, February 8

Tuesday, February 9: Best subset selection. Forward stepwise selection. Lasso. General regularized estimators.

Readings: ISL, chapter 6

Thursday, February 11: Regularizing terms as priors. Shrinkage. "Borrowing from others". James Stein.

Readings: LSI, Chapter 1*

*available online: http://statweb.stanford.edu/~ckirby/brad/LSI/chapter1.pdf

Problem set 3 released, Friday, February 12

Week 6: Midterm and Trees and natural language processing.

Laboratory 2 due, Monday, February 15

Laboratory 3 released, Monday, February 15

Tuesday, February 16: review for midterm

Thursday, February 18: Text data. Natural language processing. Topic models. Latent Dirichlet allocation. Supervised natural language processing.

Readings: handout 4

Week 7: Trees and random forest

Problem set 3 due, Monday, February 22

Tuesday, February 23: Classification trees and competitors.

Readings: ISL, chapter 8

Thursday, February 25: Random forests. Boosting. "Learning strongly form many weak learners".

Readings: ISL, chapter 8

Problem set 4 released, Friday, February 26

Week 8: Support vector machines and other classifiers

Laboratory 3 due, Monday, March 1

Laboratory 4 released, Monday, March 1

Tuesday, March 2: Basic classifiers. Maximal margin classifiers. Fisher consistency. The kernel trick.

Readings: ISL, chapter 9

Thursday, March 4: Advanced topics. Machine Learning in public policy applications. Inference with SVM. Multicategory SVM.

Week 9: Unsupervised learning and high-dimensional causal inference

Problem set 4 due, Friday, March 8

Tuesday, March 9: Unsupervised learning. Clustering. Principal component analysis. Nearest neighbors.

Readings: ISL, chapter 10

Thursday, March 11: High-dimensional causal inference.

Readings: High-Dimensional Methods and Inference on Structural and Treatment Effects (Belloni et al. 2014), Algorithmic Fairness (Kleinberg et al., 2018)

Laboratory 4 due, Monday, March 15