

# Machine Learning, Spring 2023, PPHA 30545

Professor: Guillaume Pouliot

Email : [guillaumepouliot@uchicago.edu](mailto:guillaumepouliot@uchicago.edu)

Group 1: Monday-Wednesday, 3:00-4:20

Group 2: Monday-Wednesday, 4:30-5:50

TAs:

Diego Escobar, [descobarsalce@uchicago.edu](mailto:descobarsalce@uchicago.edu)

Pavan Kumar Prathuru, [pavanprathuru@uchicago.edu](mailto:pavanprathuru@uchicago.edu)

Instructor OH: TBA

TA OH: TBA

Section 1: TBA

Section 2: TBA

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.

**No laptop policy:** Students may not use their laptops in class.

**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:

There are four “standard” problem sets and four “public policy labs”. Both count towards the Problem Sets section of the grade.

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. Not using laptops in class also counts towards participation.

**Outline:**

**Week 1:** Introduction and key concepts of statistical machine learning

*Monday, March 20:* Introduction. Course overview. The key concepts of statistical machine learning.

Readings: ISL, chapter 1

*Wednesday, March 22:* Review of basic material. Regression and hypothesis testing. The ubiquity and intuitions of multiple hypothesis testing.

Readings: ISL, chapter 2

Problem set 1 released, Friday, March 24

## **Weeks 2:** Multivariate linear regression

Laboratory 1 released, Monday, March 27

*Monday, March 27:* Factorial models. Nonlinear transformations and interactions. Matrix notation. Geometry of least-squares.

Readings: ISL, chapter 3

*Wednesday, March 29:* The regression function. Model misspecification. Best linear predictor. Identification and forecasting with singular design matrices.

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

Readings: ISL, chapter 3 Problem set 1 due, Monday, April 3

*Monday, April 3:* Multiple hypothesis testing methods. Model Selection as multiple hypothesis testing. Uniformly valid inference.

*Wednesday, April 5:* False discovery rates and the Benjamini-Hochberg theorem. ROC curves.

Problem set 2 released, Friday, April 7

Laboratory 1 due, Monday, April 10

Laboratory 2 released, Monday, April 10

*Monday, April 10:* The bootstrap, cross-validation and permutation tests.

*Wednesday, April 12:* Advanced topics. Improvements on the bootstrap. Limits of the bootstrap.

End of material for the midterm.

## **Weeks 5:** Priors, shrinkage, and regularization

Problem set 2 due, Monday, April 17

*Monday, April 17:* Best subset selection. Forward stepwise selection. Lasso. General regularized estimators.

Readings: ISL, chapter 6

*Wednesday, April 19:* 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, April 21

**Week 6:** Midterm and Trees and natural language processing.

Laboratory 2 due, Monday, April 24

Laboratory 3 released, Monday, April 24

*Monday, April 24:* midterm

*Wednesday, April 26:* Text data. Natural language processing. Topic models. Latent Dirichlet allocation. Supervised natural language processing.

Reading: handout

**Week 7:** Trees and random forest

Problem set 3 due, Monday, May 1

*Monday, May 1:* Classification trees and competitors. Readings: ISL, chapter 8

*Wednesday, May 3:* Random forests. Boosting. “Learning strongly from many weak learners”.

Readings: ISL, chapter 8

Problem set 4 released, Friday, February 25

**Week 8:** Support vector machines and other classifiers

Laboratory 3 due, Monday, May 8

Laboratory 4 released, Monday, May 8

*Monday, May 8:* Basic classifiers. Maximal margin classifiers. Fisher consistency. The kernel trick.

Readings: ISL, chapter 9

*Wednesday, May 10:* 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, Monday, May 15

*Monday, May 15:* Unsupervised learning. Clustering. Principal component analysis. Nearest neighbors.

Readings: ISL, chapter 10

*Wednesday, May 17:* Final exam

Laboratory 4 due, Friday, March 19

Time allowing, we will cover as additional material: 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)