Machine Learning in Political Sciene

POLI_SCI 490
Winter 2026

Instructor: Gustavo Diaz (gustavo.diaz@northwestern.edu)
Class: Thursdays 2:00 – 4:50pm, Scott Hall 212
Meetings: Schedule appointment (cal.com/gustavodiaz/)
Office: Scott Hall 103

Overview

This course serves as a graduate-level introduction to statistical/machine learning. It will cover common techniques to process and analyze large, often structured, data. The primary emphasis will be on methods for supervised learning, parametric methods, non-parametric methods, tree-based methods, neural networks, and ensembles. We will then briefly cover techniques for unsupervised learning, such as clustering and topic models. We will conclude with a discussion of machine learning methods for statistical inference and extensions based on classroom interest.

Content

Week	Dates	Topic
1	January 8	Introduction
2	January 15	Regression
3	January 22	Classification
4	January 29	Validation, selection, shrinkage, regularization
5	February 5	Tree-based methods
6	February 12	Neural networks
7	February 19	Ensemble learning
8	February 26	Unsupervised learning
9	March 5	Inference
10	March 12	Extensions/make-up

Learning Objectives

The goal of statistical/machine learning is prediction and description, especially in the context of large, perhaps unstructured, data sets.

By the end of the course, you will be able to:

1. Choose and apply algorithms to data relevant to your research interests
2. Evaluate the appropriateness of machine learning applications in academia, government, industry, and civil society
3. Continue learning about cutting-edge developments in the field on your own

Prerequisites

Formally, students should have taken POLI_SCI 403 (Probability and Statistics), POLI_SCI 405 (Linear Models), or equivalent familiarity with these topics and their application in statistical software like `R`, `Python`, or `Julia`. Students who do not meet these requirements should seek express approval from the instructor to take this class.

Informally, I expect willingness to learn new, complex material in a semi-supervised manner.[1] Given the breadth of the topics covered in this class, instruction cannot go much further than a mere introduction, so it is up to students to delve deeper on topics relevant to their respective research interests.

The usual norms of a graduate seminar also apply. You should participate actively, productively, and respectfully in our meetings. We are jointly responsible for creating an environment conducive to learning and discovery.

Requirements

Reading

There is no single required textbook, but we will draw heavily from:

Textbooks

- **ESL:** Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2017. *The Elements of Statistical Learning*. Springer
- **ISL:** James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021/2023. *An Introduction to Statistical Learning*. Springer
- **TM**: Kuhn, Max and Julia Silge. 2023. *Tidy Modeling with R*. O'Reilly

The initials at the beginning denote how we will refer to each book throughout the course materials. These are all textbooks that are **FREE** to read or download online. ISL has editions

[1] The joke here is that semi-supervised learning exists but it is a bit of a misnomer.

with code in `R` and `Python`, the content is the same otherwise, you should choose whichever suits your goals.

Additional readings will be linked in the Schedule section below and are usually available through Northwestern library subscriptions.

Computing

We will use R and RStudio to work on assignments and classroom demonstrations. The advantage of R is that it is free and open source, meaning that you will be able to apply everything you learn in this course anywhere else. The disadvantage is a somewhat steep learning curve. I believe the investment is worthwhile for anyone working with data or in data-adjacent careers. You are welcome to use different software for statistical computing, but I cannot guarantee I will be able to help with troubleshooting.

You can install R and RStudio on your personal computer, which is the preferred use case. You can use this link for installation instructions on Windows and MacOS. See this link for installation instructions on Chromebooks, which is a bit more involved.

You can also use Posit Cloud to access RStudio from any web browser. A free account should suffice for the purposes of this course and has the advantage of letting you access your work across devices.

If you ever need more computing resources than what a personal computer or a free Posit Cloud account allow, you should consider requesting access to the Quest Analytic Nodes from Northwestern IT. I do not anticipate this to be relevant for this course, but it may be useful in the future.

Evaluation

Your final grade in this course will depend on the following:

Participation

This course does not formally require attendance, but I do expect the usual level of accountability required in a graduate seminar. That means attending class, doing the reading, asking questions, and working to foster a productive learning environment for everyone. Your participation inside and outside the classroom will be marked as satisfactory or unsatisfactory by the end of the quarter. If your participation leans toward an unsatisfactory mark, I will notify you by the end of Week 6 and give you feedback on how to improve.

Exploration assignments

Data paper

Grading

This course uses a labor-based grading agreement, commonly known as contract grading. In this course, instead of being given a final grade based on how "good" your submitted assignments are, your final grade will be based on the amount of labor you put into the course. The goal is to decouple grades from performance and emphasize learning and effort.

You will get a default grade if you meet the contract. It will go lower if you miss parts of the contract, it will go higher if you meet the baseline plus other criteria.

To meet the baseline grading contract (B+), you should:

To get an A-, you should…

To get an A, you should…

Your grade will go below a B+ if you miss work. Unless otherwise agreed upon in writing on a case-by-case basis, the following criteria outlines how deviating from the baseline contract will impact your grade (assuming everything else constant):

[HERE]

If many of the criteria for missed work apply, you will get the lowest grade applicable.

By signing up for this course, you accept the terms of the grading contract. We will discuss potential amendments in Week 1. Amendments to the grading contract beyond this point should be agreed upon unanimously by students and the instructional team.

Northwestern University Syllabus Standards

This course follows the Northwestern University Syllabus Standards. Students are responsible for familiarizing themselves with this information.

Use of Generative AI Systems

The use of generative artificial intelligence in this course is encouraged as long as it is used to amplify humans instead of replacing them. Any form of cheating, including improper use of content generated by artificial intelligence, constitutes a violation of Northwestern's academic integrity policy.

Copilot is the University's supported artificial intelligence service. When using Copilot while actively logged in with a Northwestern account, data is stored securely in Northwestern's Microsoft tenant, and Microsoft will not use it for product improvement or to train its AI models.

Schedule

Week 1 (January 8): Introduction

Week 2 (January 15): Regression

Week 3 (January 22): Classification

Week 4 (January 29): Validation, selection, shrinkage, regularization

Week 5 (February 5): Tree-based methods

Week 6 (February 12): Neural networks

Week 7 (February 19): Ensemble learning

Week 8 (February 26): Unsupervised learning

Week 9 (March 5): Inference

Week 10 (March 12): Extensions/make-up class

No new material since it is WCAS reading week