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
Canvas: TBD
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 unstructured, data. The primary emphasis will be on methods for supervised learning, like 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	Date	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, especially in the context of large, sometimes unlabeled or 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 or professional 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 405 (Linear Models) or have equivalent familiarity with regresion models for statistical inference and their application in `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 beyond introduction, so it is up to students to delve deeper on topics relevant to their respective research interests.

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 **FREE** to read or download online. ISL has editions 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, usually available through Northwestern library subscriptions.

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

Computing

Instruction will use primarily `R` and `RStudio`, with some nods to `Python` where appropriate.

The expectation is that students will come to the first day of class with a fresh local installation of `R` and `RStudio` (or their preferred software). Please reach out to the instructor if you need assistance on this front.

Evaluation

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

1. Participation
2. Explorations (3)
3. Collaborative note taking contributions (number TBD on enrollment)
4. Final project: Data and methods paper

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.

Explorations

You will complete three exploration assignments. These are open-ended coding assignments in which you seek to accomplish two goals:

1. Teach yourself a new method or tool relevant to your research
2. Progress on your final paper

Ideally, you would start working on a data set or topic that aligns with your final paper, but this is not a firm requirement.

Explorations should be submitted as PDF files on Canvas and are due at 11:59PM on the following dates:

- Tuesday, January 27
- Tuesday, February 10
- Tuesday, March 3

Each exploration will be marked as satisfactory, unsatisfactory, or failed. You can redo any exploration marked as unsatisfactory, to which I will give detailed feedback, before the final paper deadline.

We will devote some of our class meeting time over the course of the quarter to discuss exploration topics.

Class notes

Every week, a number of students (based on enrollment) will serve as note-taking leaders. Note taking leaders will prepare summaries, discussion, and talking points for the assigned material for the week and incorporate them in a collaborative document (format TBD).

I will randomly assign topics during at the beginning of the quarter. Students are welcome to trade their turns as long as they notify the instructor at least one week ahead.

Notes will be marked as satisfactory, unsatisfactory, or failed. Unsatisfactory marks can be resubmitted at any point before the final paper deadline. Notes are due on Tuesdays at 11:59PM **before** the corresponding class meeting.

Data and methods paper

As a final project, you will write a data and methods paper. This will take the form of a research note in which you use the methods and tools covered in class to "solve" a methodological issue in your own research. The ideal project would advance an aspect of your dissertation. For example, you may write a paper about choosing the best possible classification method to code a variable from unstructured data, or use machine learning methods to estimate heterogeneous treatment effects in an original experiment.

You are required to schedule a meeting with me to discuss the topic of your replication paper. If we do not meet to discuss your paper at least once, your paper will be marked as failed.

Collaboration with people inside or outside the class is allowed with instructor approval.

Final papers are due on TBD at 9AM. You should submit your paper in PDF format through Canvas. Your paper should follow the format of a research note, up to 4,000 words. You are also required to submit a link to a GitHub repository (or equivalent) including all the code used to reproduce your tables, figures, and calculations. You may also include less important details in the appendix to keep the paper concise. You can read more about research notes in political science here.

Your final paper will be marked as *outstanding*, *satisfactory*, *unsatisfactory*, or *failed*. You are also welcome to resubmit a final paper marked as unsatisfactory at any point, but that means I will not be able to update your grade until after the final grade report deadline. I encourage submitting an incomplete paper over asking for an incomplete course grade.

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 (A-), you should:

- Be late (by a maximum of 24 hours) on no more than one assignment
- Submit the final paper before the deadline
- Have a satisfactory participation status by the end of the semester
- Complete three explorations with a satisfactory mark
- Contribute satisfactorily to the class notes every time it is your turn
- Receive a satisfactory or outstanding mark in the final paper

To get an A, you should complete the requirements listed above and also receive an outstanding mark in the final paper.

Your grade will decrease 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):

- Participation marked as unsatisfactory: One letter grade deduction (e.g. from A- to B-)
- Exploration or note-taking marked as unsatisfactory or failed: Half letter grade deduction each (e.g. from A to A-)
- Final paper marked as failed: F

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

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

1. TM Chapters 1-3
2. Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures with comments and a rejoinder by the author)." *Statistical Science* 16 (3): 199-231
3. Daoud, Adel and Devdatt Dubhashi. 2020. "Statistical modeling: the three cultures."

Week 2 (January 15): Regression

4. ISL Chapters 2-3
5. ESL Chapters 2-3.2
6. TM Chapters 4-8

Week 3 (January 22): Classification

7. ISL Chapters 4, 9
8. ESL Chapter 4, 12, 13

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

9. ISL Chapters 5-6
10. ESL Chapter 3.3-3.9, 7
11. TM Chapters 9-15

Week 5 (February 5): Tree-based methods

12. ISL Chapter 8
13. ESL Chapters 9-10, 15

Week 6 (February 12): Neural networks

14. ISL Chapter 10
15. ESL Chapter 11

Week 7 (February 19): Ensemble learning

16. ESL Chapter 16

Week 8 (February 26): Unsupervised learning

17. ISL Chapter 12
18. ESL Chapter 14

Week 9 (March 5): Inference

19. Fuller, Sam and Jack T. Rametta. 2024. "Causal Forest and Double Robust Machine Learning for Political Science."

20. Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogenous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228-1242

21. Chernozhukov, Victor et al. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1): C1-C68

22. Ratkovic, Marc. 2023. "Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression.". *American Political Science Review* 117 (3): 1053-1069

    - Video version

23. Egami, Naoki, Musashi Hinck, Brandon M. Steward, and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses."

24. Morucci, Marco and Arthur Spirling. 2024. "Model Complexity for Supervised Learning: Why Simple Models Almost Always Work Best, And Why It Matters for Applied Research."

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

No new material since it is WCAS reading week