Data Mining and Machine Learning
STSCI 3740/5740
Fall 2024

Course personnel

Instructor

- Dr. Nayel Bettache, Visiting Assistant Professor in the Dept. of SDS
  Office: Academic Surge Facility B 158 220 Tower Road, Ithaca, NY 14853
  E-mail: nbb45@cornell.edu
  Office Hours: Tuesday, 2:30 PM - 3:30 PM, and by appointment.

Teaching Assistants

- Yixin Shen, Ph.D. student in the Dept. of SDS
  E-mail: ys964@cornell.edu
  Office hours: TBA
- Minjie Jia, Ph.D. student in the Dept. of SDS
  E-mail : mj598@cornell.edu
  Office hours: TBA

Course Description

This course provides an introduction to the fundamental concepts and techniques in statistical learning and machine learning, with a focus on understanding the theoretical underpinnings of various machine learning algorithms and their implementation in R (and tentatively in Python).

Course Objectives

By the end of this course, students will be able to:

1. Explain the concepts of regression, classification, and clustering, and apply them to real-world problems.
2. Implement machine learning algorithms in R (and tentatively in Python).
3. Evaluate and compare the performance of different machine learning models.
4. Understand the trade-offs involved in model selection and regularization.

Lecture meeting times

Lecture (Professor): Tues &Thurs 11:40am-12:55pm (Phillips Hall 101)

Prerequisites and textbooks

- Prerequisites: **CS 1112, MATH 2220, STSCI 3200, STSCI 3080 or MATH 4710** or equivalents. Students must have a good command of basic statistics, probability, linear algebra and calculus. Proficiency in R or Python programming or willingness to learn is required.
- Required textbook: **An Introduction to Statistical Learning with Applications in R**, 2nd Edition, Springer, 2021 by G.James, D.Witten, T.Hastie, and R.Tibshirani. (ISLR) `https://www.statlearning.com`.
- Supplementary textbook: **The Elements of Statistical Learning**, 2nd Edition, Springer, 2009 by T.Hastie, R.Tibshirani, and J.Friedman. (ESL) `https://hastie.su.domains/Papers/ESLII.pdf`.

Course Schedule

The following schedule is a general outline that we plan to follow. Depending on the pace of the course, some topics may be explored in greater detail, while others might be adjusted or omitted. Assignments are currently planned to be released on Thursdays of the corresponding week, though this is subject to change.

Week 1: 8/26

- Topic: Introduction

Week 2: 9/02

- Topic: Statistical Learning

Week 3: 9/09

- Topic: Linear Regression

Week 4: 9/16

- Topic: Linear Regression

Week 5: 9/23

- Topic: Classification
- Assignment: Homework 1

Week 6: 9/30

- Topic: Classification
- Assignment: Midterm 1

Week 7: 10/07

- Topic: Resampling Methods
- Assignment: Homework 2

Week 8: 10/14 (half week)

- Topic: Linear Model Selection and Regularization

Week 9: 10/21

- Topic: Linear Model Selection and Regularization
- Assignment: Homework 3

Week 10: 10/28

- Topic: Linear Model Selection and Regularization
- Assignment: Midterm 2

Week 11: 11/04

- Topic: Nonlinear Models

- Assignment: Homework 4

Week 12: 11/11

- Topic: Tree-Based Methods

Week 13: 11/18

- Topic: Support Vector Machine
- Assignment: Homework 5

Week 14: 11/25 (half week)

- Topic: Deep Learning

Week 15: 12/02

- Topic: Unsupervised Learning

Course materials

The materials for this class will be uploaded on `https://www.nayelbettache.github.io`. All materials listed below will be available online, at this site. It is entirely your responsibility to download them as needed. A brief description of these materials follows.

- The syllabus should be used as a reference throughout the year for important dates, including exams, and for course policies.
- Lecture notes will be posted on this website as the semester progresses. While these notes form the foundation of my lectures, additional insights and details will be provided during class.
- The lecture notes are designed to complement, not replace, the textbook. Their purpose is to guide you through new material more easily. It is your responsibility to thoroughly read both the notes and the corresponding textbook chapters. Ensure you identify the relevant sections and subsections in the textbook that align with the lecture notes. This constitutes your reading assignment for the semester.

Homeworks, exam schedule and grading policy.

Your grade in this class will be based on homeworks and exams, as below.

1. **Homeworks.** You will receive five assignments counting towards **20 %** of the grade. The lowest homework score will be dropped, with the remaining four assignments weighted equally. For students in 5740, you may encounter one or two additional questions per assignment, which are required for 5740 but optional for 4740 (offering bonus points for 4740 students). Late homework submissions will incur a 20% penalty if submitted within 24 hours past the deadline; submissions beyond that will not be accepted. Solutions will be posted on the course website two days after the submission window closes. Please refer to the Course Schedule above for deadlines.

2. **Midterms.** There will be two in-class tests, each during regular lecture times, collectively accounting for **50%** of your final grade. Both midterms will carry equal weight, and the schedule is provided in the Course Schedule above.
   Each test will cover the material discussed in class up to the exam date, including problems solved in lectures and all the homeworks due before the exam. I will provide an overview of the exam in class and post a detailed outline of the required materials before each test. All exams are closed-book, and the use of any electronic devices is strictly prohibited. This includes computers, calculators, cellphones, and other electronic gadgets.
   Students with approved extended time: please see the section on accommodations below.

3. **Final project.** The final project for this course will be a take-home data analysis assignment, designed to be completed at the end of the semester. The project will require students to work in groups, and the datasets along with specific questions for analysis will be distributed around October 20. Students are expected to form groups of 3 to 4 members. These groups should be finalized and approved by the instructor no later than November 1. Any students who have not joined a group by this deadline will be assigned to a group by the instructor.

   The final report, which documents the results of your analysis, must be

submitted as a PDF file by December 16. If the report is submitted late, a 20% penalty will be applied if it is received within 24 hours after the deadline; reports submitted after this period will not be accepted. The report should be no longer than 8 pages, formatted in a standard style with a font size of 12. It should demonstrate the application of appropriate methods discussed throughout the course, present findings clearly, and provide accurate interpretations of the results.

All data analysis must be conducted using R or Python, and the scripts used in your analysis must be submitted alongside the report, although these scripts will not count towards the 8-page limit. Your project will be graded based on the effective application of the appropriate methods, the clarity and organization of the report, the accuracy of the interpretations, and the reproducibility of your analysis using the provided scripts.

This project is an integral part of the course and will allow you to apply the knowledge and skills you have developed throughout the semester in a practical and meaningful way. It is an opportunity to demonstrate your understanding of the course material and your ability to conduct and present a thorough data analysis.
There is no curving of grades in this class. Your final grade will be based entirely on your performance.

The correspondence between your final percentage and your final letter grade is as follows:

- 97 or more: A+; $[93, 97)$ A; $[90, 93)$ A-; $[86, 90)$ B+
- $[83, 86)$ B; $[80, 83)$ B-; $[76, 80)$ C+; $[73, 76)$ C; $[70, 73)$ C-
- $[66, 70)$ D+; $[63, 66)$ D; $[60, 63)$ D-; Less than 60: F.

Students with disabilities and exam accommodation

- **Students with Disabilities**
  Students with disabilities are encouraged to engage fully in this course, and your access needs are a priority. To ensure that your approved accommodations are arranged in a timely manner, you must request your accommodation letter via the SDS Student Portal by September 23.

For students who are already registered with the Student Disability Services (SDS), please note that once you request your accommodation letter, it may take up to 48 hours for the letter to be processed and sent to me. If you are not yet registered with SDS, be aware that the process to register and receive new accommodations can take up to three weeks. Once approved, you will be able to request your accommodation letter for this course.

If you are approved for accommodations later in the semester, it is important that you request your accommodation letter as soon as possible to avoid any delays in receiving the necessary support.

- **Students with Exam Accommodations**
  Regarding exam accommodations, this course is participating in the Alternative Testing Program (ATP). All exams will be centrally managed by the ATP, and relevant information will be communicated through SDS-testing@cornell.edu and your SDS Student Portal. It is important to stay informed by reading these communications and visiting sds.cornell.edu/atp for additional details about the ATP process.

  Starting in Fall 2023, students no longer need to request each individual exam. However, if you have an academic conflict with a scheduled exam time, you must submit an "exam request form" in the SDS Student Portal. All requests for conflict exams must be submitted no later than 10 business days prior to the exam date, and conflict exams will be scheduled at standard times.

  For all relevant information and to manage your accommodations, please visit the SDS Student Portal at sds.cornell.edu.

Academic Integrity

Course materials provided in this class are the intellectual property of the instructor. Students are strictly prohibited from buying, selling, or distributing any course materials without the express permission of the instructor. Engaging in such unauthorized activities is considered academic misconduct and will be treated accordingly.

Every student in this course is expected to adhere to the Cornell University Code of Academic Integrity. All work submitted for academic credit must be

the student's own original work. The use of AI resources, including tools like ChatGPT, is strictly prohibited in this class.

Wellness Resources

The material provided below has been thoughtfully compiled by students from the Body Positive Cornell organization. It offers a well-researched and comprehensive list of well-being resources available on campus. For detailed information and guidance, please refer to the following resource:

- Mental Health Resources Guide 2022-23
- Cornell Wellbeing Resources