APSTA-GE 2047: Messy Data and Machine Learning

Class location and time
The class will meet on Thursdays from 10:15am-12:15pm (Eastern) in 194 Mercer, Room 208.

Contact information
Instructor: Ravi Shroff (ravi.shroff@nyu.edu)
Student Hours: Tue, 3pm-4pm Eastern in person (Kimball Hall #202W) and simultaneously on zoom (link available on Brightspace), or by appointment

Course Assistant: Angela He (dh2551@nyu.edu)
Student Hours: Wed, 8pm-9pm Eastern on zoom (link available on Brightspace), or by appointment

NYU values an inclusive and equitable environment for all students. I hope to foster a sense of community in this class and consider it a place where individuals of all backgrounds, beliefs, ethnicities, national origins, gender identities, sexual orientations, religious and political affiliations, and abilities will be treated with respect. It is my intent that all students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. If this standard is not being upheld, please feel free to speak with me.

Course Description
This course exposes students to the complex real-world datasets commonly used in machine learning applications. The course provides an accessible introduction to supervised machine learning, while covering aspects of data collection and cleaning. Specific topics include model construction, evaluation, and regularization, as well as web scraping, text data, feature construction, and measurement error. Students complete homework sets and a final project.

Disclaimer: since this is a first course in machine learning, there is a tremendous amount of material that we will *not* cover, including many standard algorithms for supervised learning (deep learning methods, support vector machines, etc.). With one exception, we will not cover unsupervised learning methods. Since the focus of this course is on practice, we will largely avoid mathematical theory in favor of building intuition.

Course Prerequisites
The main prerequisite is an intermediate course in statistics at the level of APSTA-GE 2003. It will be assumed that students have basic facility with R (e.g., loading and manipulating data, and familiarity with core functions such as lm and plot) at the level of APSTA-GE 2352 (which may be taken concurrently), however any additional experience will be helpful. Prior knowledge of classification and clustering at the level of APSTA-GE 2011 will also be helpful, although it is not required.

Learning Objectives

Upon completing this course, students should be able to:

1) Identify, collect, and clean a complex real-world dataset.
2) Implement and evaluate commonly used supervised learning techniques, and clearly explain how these techniques work.
3) Perform data cleaning and machine learning techniques in R in a robust and reproducible manner.
4) Assess and articulate ethical and other concerns and limitations associated with using specific datasets for supervised learning.

Course Format

We will meet in person for 120 minutes per week with a short break in the middle; each meeting may include both lectures and in-class group work. Lecture slides and course materials will be available on the course website on Brightspace.

Course Grading

The grade for this course will be based on:

1. Participation **(10%)**. Please do your best to be engaged! You can show your engagement in the following ways (*please keep track of how you participated, as I will ask for this at the end of the semester to help determine your grade*):
    a. Attending lecture, and asking and answering questions either during lecture, on the Brightspace forum, or on the slack channel for this class;
    b. Completing in-class group work;
    c. Attending student hours; or by attending seminars or talks related to the subject material, especially if you ask questions.

    Do not use your phone during class, and limit the use of laptops/tablets except as instructed.
2. Individual homework assignments **(60%).** Assignments reinforce and extend understanding of concepts covered in class. These assignments will include: questions that assess conceptual understanding, hands-on data manipulation tasks, and implementation of machine learning techniques. You will be expected to clearly and concisely explain your results in writing. *Doing the homework is essential for succeeding in this class and learning the material!*
3. A final group project **(30%)**. This project will synthesize all the skills you'll learn in the course: problem identification, dataset collection and cleaning, and algorithm implementation. You will write a short paper explaining your work and results.

Grading Scheme

Specific grading criteria for each type of assignment above will be provided. The following thresholds indicate a numerical grade *sufficient* to guarantee the corresponding letter grade. However, final letter grades assigned will depend on the distribution of final numeric grades in the class (in particular, I may grade more leniently than this rubric suggests, but I won't grade more harshly):

93%: A
90%: A-
87%: B+
83%: B
80%: B-
77%: C+
73%: C
70%: C-
67%: D+
63%: D
60%: D-
<60%: F

Deadlines and Homework Policies
Homework assignments are to be submitted on Brightspace *before* the beginning of the lecture (10:15am Eastern time) on the due date. Late problem sets will be reduced according to the following schedule:

- If turned in **within 24 hours** of the deadline, 5 percentage points will be deducted from the problem set grade.
- If turned in **after 24 hours, but within 4 days** of the deadline, 10 percentage points will be deducted from the problem set grade.
- If turned in **over 4 days** after the deadline, no credit will be given for the problem set.

If you fall ill, you do not need to provide additional information or documentation to the instructor, and you will be given an extension corresponding to the number of days you were ill. Barring illness, exceptions to the policy above are granted only in extreme circumstances and require written documentation. Examples include a documented accommodation approved by the Moses Center for Student Accessibility or hospitalization. Poor time management does not count as an exceptional circumstance.

In general, “A” work requires: correct numerical answers, correct implementations of methods, accurate and understandable interpretations of any results, and clear explanations regarding the connections of any results to the scientific or real-world phenomenon being examined. All work is to be submitted on Brightspace and should be professional in appearance; in particular, sentences should be complete, and answers must be clear and coherent.

You should work on the problem sets by yourself. You may discuss problems with your classmates, the course assistant, and the professor, but **your submitted work must be your own** (please see the Academic Integrity statement here). In particular, you may **not** copy code from your classmates; this will be deemed plagiarism and sanctioned according to NYU guidelines. You may **not** use ChatGPT or similar generative tools when doing your homework unless explicitly instructed to do so.

You should work in groups of 2-3 students on the final project. Turn in one report per group, and indicate group members; you may have the opportunity to assess the contributions of other group members, and individual grades may be assigned. Your project must have a clearly stated and motivated goal or research question, and must use one or more applicable datasets. You must apply at least one supervised learning method and write a report (~10 pages) summarizing related literature, describing methods, detailing results and plots, and discussing the implications of your findings. *By Week 7, you will submit a short description of your planned project, which will include the substantive question that you will examine, a brief summary of the data source(s) you plan to use, and an outline of the methods you will potentially apply.*

The final project will be graded according to the following rubric (/**100**):

1) Clearly stated and motivated goal/research question: **20 pts**;
2) Cleaning/compilation and description of data (utilizing at least one non-Kaggle dataset): **20 pts**;
3) Description and use of at least one ML method; the connection between the problem and methodology should be clear and compelling: **20 pts**;
4) Description of results and implications, including limitations: **20 pts**;
5) Organization, clarity, and correctness of writing and plots: **15 pts**;
6) Literature review: **5 pts**.

You must also submit all your code, which should be clear and easy to follow (I recommend including a readme file).

Attendance policies:

You are expected to attend class in person unless you are feeling unwell. If you anticipate missing class due to sickness, **please let me know in advance**. Participation grades will *not* be penalized for missing class due to illness if you inform me about your absence in advance, and HW extensions (for illness, if requested) will be granted as appropriate. Also:

1) Lectures will be recorded, but will only be made available to students who either attended lecture or who have an excused absence due to, e.g., illness.
2) Students are not permitted to attend class remotely.
3) If *I* get sick, we may temporarily switch to remote instruction, but hopefully not for more than one week. If this is necessary, I will either hold class remotely over zoom at our usual time (links will be posted on Brightspace) or, if I'm unable to hold class at all, I'll record the lecture for you to watch later.

Required Readings/Text

[ISL] An Introduction to Statistical Learning - James/Witten/Hastie/Tibshirani (free)
[R4DS] R for Data Science - Grolemund, Wickham (free)
[PG] Pro Git - Chacon, Straub (free)
[TMWR] Text Mining with R - Silge, Robinson (free)

Recommended Reading

1. RMarkdown: the definitive guide (Xie, Allaire, Grolemund)

2. "The task is a quantum of workflow" (Ball 2016)
3. "Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy" (Goel, Rao, Shroff 2016)
4. "Quantitative Analysis of Culture Using Millions of Digitized Books" (Michel, Shen, Aiden, et al. 2011)
5. "Web-scale pharmacovigilance: listening to signals from the crowd" (White, Tatonetti, Shah, et al. 2013)
6. Racial disparities in automated speech recognition (Koenecke, Nam, Lake, et al. 2020)
7. Dissecting racial bias in an algorithm used to manage the health of populations (Obermeyer, Powers, Vogeli, et al. 2019)
8. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Buolamwini, Gebru 2018)

Course Outline [this may change slightly over the course of the semester]:

Date	**Subject**	**Topics**	**Reading**
Week 1 1/26/2023	Introduction (**HW 1 assigned**)	Inference vs. prediction; parametric vs. nonparametric methods; a global view of ML; examples of datasets	ISL Chs. 1, 2.1 R4DS Chs. 1, 4, 6, 8
Week 2 2/2/2023	Manipulating data; workflow	Introduction to dplyr	R4DS Chs. 5, 9, 10, 11, 12, 13, 17, 18
Week 3 2/9/2023	Reproducibility (**HW 1 due/HW 2 assigned**)	Version control; introduction to the terminal; organizing files	PG Ch. 1, Ch. 2.1-2.4, (Ball 2016)
Week 4 2/16/2023	Introduction to statistical learning and classification	Accuracy vs. interpretability; overfitting; KNN	ISL Ch. 2.2
Week 5 2/23/2023	More on classification **(HW 2 due/HW 3 assigned**)	Logistic regression; generative models	ISL Ch. 4.1 - 4.3
Week 6 3/2/2023	Model evaluation	Model accuracy continued; AUC; precision and recall; calibration	ISL Ch. 4.4.3 (Goel, Rao, Shroff 2016)
Week 7 3/9/2023	Model selection and assessment (**HW 3 due/HW 4 assigned**) (**Project plan due**)	Cross-validation; the bootstrap	ISL Ch. 5
No class on 3/16 - Spring Break!			
Week 8 3/23/2023	Web scraping	APIs;	Instructor notes
Week 9	Text data	cleaning; lemmatization; bag-of-words;	TMWR Ch.1,3,4,5

3/30/2023	(**HW 4 due/HW 5 assigned**)	tf-idf;	(Michel et al. 2011) (White et al. 2013)
Week 10 4/6/2023	Feature engineering	Interactions; features in text data; topic modeling;	TMWR Ch. 6,
Week 11 4/13/2023	Subset selection and regularization (**HW 5 due/HW 6 assigned**)	Subset selection; regularization: ridge regression and lasso; high dimensionality	ISL Ch. 6
Week 12 4/20/2023	Tree-based methods	Decision trees; bagging; random forests; boosting	ISL Ch. 8
Week 13 4/27/2023	Algorithmic Fairness (**HW 6 due**)	Examples from gender classification, healthcare, and speech-to-text	(Koenecke et al. 2020) (Obermeyer et al. 2019) (Buolamwini, Gebru 2018)
Week 14 5/4/2023	NO CLASS (**Final project due**)		

Academic Integrity:
All students are responsible for understanding and complying with the New York University's policies on academic integrity. A copy of this policy for the Steinhardt school is available here.

Access and Accommodations:
New York University is committed to ensuring equal educational opportunity and accommodations for all students. Students should contact the Moses Center for Student Accessibility at (212) 998-4980. The Center will work with students to determine appropriate and reasonable accommodations that support equal access.