STAT 195: Introduction To Supervised Learning, Spring 2024

Instructor: Pragya Sur (pragya@fas.harvard.edu)

Lectures: Mondays and Wednesdays at 4:30-5:45 PM

Course Webpage: https://canvas.harvard.edu/courses/130033

Discussion forum: We will use Ed for discussions.

Teaching Fellow: Kuanhao Jiang (kuanhaojiang@g.harvard.edu)

Sections and Office Hours:TBD

Text: The course will draw from the textbook *Elements of Statistical Learning* (ESL) by Hastie, Tibshirani, and Friedman, freely available in pdf form from the authors online (click 'Download the book PDF' in https://web.stanford.edu/~hastie/ElemStatLearn/), as well as from recent academic papers. The latter can either be found freely online, or through Harvard Hollis.

Prerequisites: CS 181 and Stat 110 or equivalents are *required* (We allow CS/Stat 109A to count as equivalent of CS 181), Stat 111 or other substantial exposure to at least the basic ideas of statistics is *strongly recommended*, but not required.

Grading: Five homework assignments (9% each; total of 45%), course project/literature review (50%), and class participation (5%).

Homework Policies: There will be a total of five homework assignments. Typically, homeworks will be released between Wed-Friday and be due on Canvas two weeks later, making suitable adjustments so you receive two weeks to work on the homework without any other course deliverables within those two. The only exception to this will be the last homework that will receive a bit less than 2 weeks due to the semester timeline. Collaboration is allowed but students must write up their own solutions and report any collaborators when they turn in the assignment. For homeworks Generative AI is not allowed to be used in any form. Each student will have 48 cumulative hours of late time (measured on Canvas) forgiven after which assignments turned in late will receive no credit, and each student's lowest homework score will be dropped; this sentence only applies to homeworks, not any part of the course project/literature review component.

Project/Literature Review Component: In lieu of exams, we will have students work on a course project with a literature review component. For the first two parts of this component, students will work in groups to synthesize and give short presentations on recent machine learning papers. For the last part of this component, students will work individually and submit a 3-pg write up on a course project. We will have short oral exams for every student where we ask them 3 questions from their course project write-up to test their understanding. This oral component is being added this year to ensure the project was conducted by the student and not by tools such as Generative AI. This entire component accounts for 50% of the course grade and has three parts. I describe the three parts in further detail below.

In the first part (resp. second part) , each group will present an *applied* (resp. *theoretical or methodological*) machine learning paper; these presentations will be held during class times. The exact weeks when these presentations occur will be determined later. A curated list of papers to choose from will be provided, although students are welcome and encouraged to choose a paper on their own (subject to instructor approval). An early lecture will be devoted to guidance and best practices for reading and presenting research papers. Each group will meet with the instructor during office hours before these presentations to discuss the paper. We seek to make the aforementioned presentations interactive, and students will be asked to raise questions from fellow students' presentations. At the end of the presentation sessions, we will compile questions from the audience, and each group will be expected to respond to these questions in writing (could be

typeset/handwritten and scanned), within the following one week. Details regarding the format of questioning will be clarified before the presentation sessions.

In the third part of this component, each student will work individually (no longer in groups). They will take the applied machine learning paper that they presented during their course presentations and use concepts learned in class or concepts presented by themselves or a fellow student during the theory/methods presentation days to improve the approach taken in the applied paper. Using at least one theoretical insight or methodological idea from the course or theory/methods round of presentations of the course, students will be asked to explain how it can be used to positive effect in the context of their applied machine learning paper. Students may use Generative AI to revise (meaning only to revise the English) their project report. To ensure this policy is strictly followed, and that students are composing their project reports themselves without the use of Generative AI, we will hold oral exams during the exam week. These will be short exams where course staff will ask each student 3 questions based on their project reports.

Grade division will be as follows: 10% for first part presentation, 10% for second part presentation, 5% for raising questions and responding to questions, 25% for write up and performance on the oral exam.

Class Participation: Lectures will be active: I will stop a few times during each lecture and have everyone work individually for roughly 2 minutes a question I pose. Because of the importance of the active learning component of the lectures, 5% of the course grade will come from class participation: after each such 2-minute working time, I will solicit thoughts on the proposed question from the class. As long as a student regularly attends lectures (at least $\sim 80\%$ of them) and participates in the discussion following the active learning exercises, they will receive full points for class participation.

Goals: The high-level goal of the course is to introduce and prepare students for theoretical and methodological research in statistical machine learning. This will center on understanding the fundamentals of when and how different machine learning algorithms achieve high prediction accuracy. By the end of the course, students will gain an understanding of how to use what they know (both qualitatively and quantitatively) about their data in a given problem to choose the right machine learning method, and also how and when such methods can be further improved. The goal of the active learning component of the class is to enhance understanding by having students engage with lecture material while it is being taught.

Tentative Topics covered: Nearest neighbors, no free lunch theorems, curse of dimensionality, structured learning, subset selection, shrinkage methods, principal components regression, optimism, effective number of parameters, cross validation, ensembling, implicit regularization and interpolation, transfer learning, algorithmic fairness, conformal inference, robustness, causal inference using machine learning.

Generative AI policy Through the course, you are allowed to use Generative AI to deepen and broaden your understanding of the topics covered in class. In particular, if there are topics mentioned in passing in class on which modern research is performed, you may use Generative AI to find more sources to learn the corresponding material from. However, use of Generative AI for homeworks, or any part of the course project/literature review component (other than using it to revise a penultimate version of your project report to improve your English, not change the content) is not allowed.