STAT 195: Statistical Machine Learning, Fall 2020

Instructor: Lucas Janson (`ljanson@fas.harvard.edu`)

Lectures: Tuesday and Thursday at 1:30 - 2:45 PM (Eastern Time Zone) on Zoom; lectures will be recorded and posted on Canvas immediately after they finish.

Students are expected to attend all lectures live at the scheduled class time. To maintain a vibrant and engaging environment for everyone and to enable the instructor to gauge student learning in real time, students are expected to keep their camera on during class, although this will not be strictly enforced as we understand there are legitimate reasons for turning off one's camera at times.

Course Webpage: `https://canvas.harvard.edu/courses/77262`

Discussion forum: Slack workspace; found and joined on course webpage

Teaching Fellow: Alexandre Bayle (`alexandre_bayle@g.harvard.edu`)

Sections and Office Hours: If enrollment is n, there will be $\lceil n/12 \rceil$ sections and the same number of TF-held office hours, plus two more instructor-held office hours; all times will be determined by online poll in the first week.

Text: The course will draw considerably from recent academic papers to which students will be given access, as well as 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/`).

Prerequisites: CS 181 or equivalent is *required*, although I am defining "equivalent" liberally—this course will assume you have had significant exposure to and practice implementing at least some basic machine learning algorithms, so if this describes you but you haven't taken CS 181, just send me an email detailing your machine learning background so I can check (I have historically granted nearly every request to replace the CS 181 prerequisite with an "equivalent", but you must check with me). 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%), three-part literature review (15% per part; total of 45%), and class participation (10%). The course is letter-graded by default, but you may switch to SAT/UNSAT if you prefer.

Homework Policies: A total of five homework assignments will each be assigned on Thursdays and be due on Canvas two weeks later on the next-to-next Thursday at the beginning of class (see schedule). Collaboration is allowed but students must write up their own solutions and report any collaborators when they turn in the assignment. 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 parts of the course's literature review component.

Literature Review Component: In lieu of exams or a course project, a new component of the course this year will have students work in groups to synthesize and give short presentations on recent applied and theoretical machine learning papers. The first two parts will have each group prepare a short presentation on a recent academic paper; in the first part each group will present an *applied* machine learning paper and in the second part each group will present a *theoretical or methodological* machine learning paper; presentations will be held in the fifth and tenth week of class for the two parts, respectively (see schedule below). 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, and an example presentation will be given by the TF.

Each group will meet with the instructor during office hours before each presentation to discuss the paper. The third part of the literature review component applies the concepts learned in the class to understand and possibly improve the approach taken in the applied paper the student presented; this will be due at the end of the course (see schedule below).

Class Participation: Lectures will be active: I will stop a few times each lecture and send everyone into 2-person breakout rooms to discuss for roughly 2 minutes a question I pose. Because of the importance of the active learning component of the lectures, 10% of the course grade will come from class participation: during each active learning exercise, I will visit a few random breakout rooms to check in on the discussion, and as long as a student regularly attends lectures (at least $\sim 80\%$ of them) and participates in the active learning exercises, they will get full points for class participation. Students will also be asked to fill out a short 1-minute online survey at the end of each lecture so I can gauge their learning and address common points of confusion at the beginning of the next lecture; the forms will ask for students' names so I can follow up individually in certain cases if needed, but no part of the course grade will depend on these forms.

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 Schedule:

Date	Homework	Topics
9/3		Class schedule/policies, Optimal prediction functions, "no free lunch" theorems, curse of dimensionality, bias-variance tradeoff, test error, VC dimension, cross-validation.
9/8		
9/10	1 out	
9/15		
9/17		
9/22		
9/24	1 due	
9/29		*Student group presentations on applied machine learning papers*
10/1	2 out	*Student group presentations on applied machine learning papers*
10/6		Maximum likelihood, Bayes, bootstrap, bagging, model averaging, linear regression, shrinkage, ridge regression, lasso, elastic net, principal components, linear discriminant analysis, logistic regression, separating hyperplanes, splines, smoothing, local likelihood.
10/8		
10/13		
10/15	2 due, 3 out	
10/20		
10/22		
10/27		
10/29	3 due	
11/3		*Student group presentations on theoretical machine learning papers*
11/5	4 out	*Student group presentations on theoretical machine learning papers*
11/10		Trees, boosting, random forests, implicit regularization, machine learning for causal inference, interpolation, conformal inference, transfer learning, double descent, adversarial robustness.
11/12		
11/17		
11/19	4 due, 5 out	
11/26		
12/1		
12/3	5 due	
12/7		[no class] *third part of literature review component due*