STATISTICS AND DATA SCIENCE 355
INTRODUCTORY MACHINE LEARNING

Syllabus, Fall 2019

Introductory Machine Learning covers the key ideas and techniques in machine learning without the use of advanced mathematics. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods and a more formal understanding of how and why they work. Assignments give students hands-on experience with the methods on different types of data. Topics include linear regression and classification, tree-based methods, clustering, topic models, word embeddings, recurrent neural networks, sparse coding and deep learning. Examples come from a variety of sources including political speeches, archives of scientific articles, real estate listings, natural images, and several others. Programming is central to the course, and is based on the Python programming language.

Schedule

LECTURES	Tuesday and Thursday 9:00-10:15 am	Osborn Memorial Laboratories 202

Teaching Staff

Instructors

John Lafferty		`john.lafferty@yale.edu`
Office hours:	Tue 4:30–5:30	location TBA

Teaching Fellows

Parker Holzer		`parker.holzer@yale.edu`
Office hours:	Wed 4:00-5:00	location TBA
Xinyi Zhong		`xinyi.zhong@yale.edu`
Office hours:	Thu 5:00-6:00	location TBA

The course will also have Undergraduate Learning Assistants (ULAs), to be announced later.

Prerequisites

Prerequisites: At least two of the following courses: S&DS 230, 238, 240, 241 and 242; previous programming experience (e.g., R, Matlab, Python, C++), Python preferred. The course will make extensive use of Python programming, using Jupyter notebooks. A beginner's guide to installing Python and Jupyter on your computer is here: `https://bit.ly/22KVCfsV`

Textbook

The course will partly use the text "An introduction to statistical learning," by G. James, D. Witten, T. Hastie, and R. Tibshirani, Springer (2013), `http://www-bcf.usc.edu/~gareth/ISL/`

Most of the material in the course will not be based on any textbook; notes and readings will be posted on Canvas.

Course Structure and Grading

The course will have a standard lecture format. Some classes will involve working through examples in software. Students are encouraged to bring a laptop to class to work along with these examples.

Assignments will be posted roughly every 10 days (week and a half), and submitted before midnight on the due date. Assignments will typically include a mix of problem solving and data analysis (coding). Python will be the course programming language.

Two in-class mid-semester exams will be given. Each exam will involve some programming, true/false and multiple choice questions, and some short answer questions. Two short 10 minute in-class quizzes will also be given. The final exam will have a mix of programming and written questions. Each of these components will be weighted as follows to determine a final grade:

- Assignments: 50%
- Mid-semester exams: 25%
- Final exam: 20%
- Quizzes: 5%

Policy on Assignments and Collaboration

Assignments must be submitted to Canvas on the day they are due, typically at midnight. Both written and programming portions of assignments will be submitted electronically. Each student's lowest assignment grade will be dropped. Late assignments will not be accepted, unless there are extenuating circumstances (e.g. family emergencies). Undergraduate students must obtain a Dean's excuse for any late submission to be graded.

Collaboration on homework assignments with fellow students through discussion of ideas is encouraged. However, you may not share written work or code; solutions must be written by yourself. Any collaboration should be clearly acknowledged, by listing the names of the students with whom you have had any discussions concerning the assignment.

See http://ctl.yale.edu/writing/wr-instructor-resources/addressing-academic-integrity-and-plagiarism for further information.

Course Materials, Calendar and Discussion Board

Course materials will be posted on the Canvas website, and will be updated throughout the semester. *Piazza* will be used as a forum for discussion and questions. Students are strongly encouraged to participate on Canvas by both asking questions and answering other students' questions.

A preliminary schedule of topics, exams, and assignments is given below. The schedule of topics and assignments is subject to minor changes. However, the quiz and exam schedule is fixed, to allow students to make appropriate plans. Lecture slides, readings, distributed notes, and online tutorials will be posted on Canvas.

Week	Date	Topic	Out/Due
1	Aug 27	—	
	Aug 29	course introduction	
2	Sept 3	linear regression and classification	
	Sept 5		assn 1 out
3	Sept 10	stochastic gradient descent	
	Sept 12		
4	Sept 17	bias and variance, cross-validation	assn 1 in; assn 2 out
	Sept 19		
5	Sept 24	tree-based methods	
	Sept 26		assn 2 in; assn 3 out
6	Oct 1	PCA and dimension reduction	quiz 1
	Oct 3		
7	Oct 8	mixtures and Bayes	assn 3 in; assn 4 out
	Oct 10		
8	Oct 15	—	midterm exam 1
	Oct 17	no class (October recess)	
9	Oct 22	topic models	
	Oct 24		assn 4 in; assn 5 out
10	Oct 29	language models, word embeddings	
	Oct 31		quiz 2
11	Nov 5	introduction to neural networks	assn 5 in ; assn 6 out
	Nov 7		
12	Nov 12	autoencoders and multilayer networks	
	Nov 14		assn 6 in; assn 7 out
13	Nov 19	—	midterm exam 2
	Nov 21	recurrent neural networks	
14	Nov 26	no class (Thanksgiving break)	
	Nov 28		
15	Dec 3	special topics and review	assn 7 in
	Dec 5		