Statistics 154: Modern Statistical Prediction and Machine Learning

Syllabus, Spring 2017

Lectures

Time: Tue/Thu, 2:00 PM - 3:30 PM Place: 9 Lewis

Lab section

Time: Monday, 2-4 and 4-6, 330 Evans

Instructor

Noureddine El Karoui

nkaroui@berkeley.edu
Office hours: Thursday, 3:30-5:00

Graduate student instructor

Raj Agrawal (r.agrawal@berkeley.edu)

Office hours: TBD

Texts

Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., 5th printing.

Available as a free pdf download: see here or at Springer Link (where you can order a softcover version for around $25)

http://statweb.stanford.edu/~tibs/ElemStatLearn/

http://link.springer.com/book/10.1007/978-0-387-84858-7/page/1

Optional: Peter Dalgaard, *Introductory Statistics with R*, 2nd ed. Available as a pdf through SpringerLink: see here.

http://link.springer.com/book/10.1007/978-0-387-79054-1/page/1

Prerequisites

- A semester of multivariate calculus or the equivalent, esp. partial derivatives; e.g., Math 53

• A semester of linear algebra or the equivalent (matrices, vector spaces); e.g., Math 54; Math 110 strongly recommended. We will use eigenvalues extensively.

• A semester of statistical inference or the equivalent; e.g., Stat 135

These are "real" prerequisites: taking the course without them would be a frustrating experience.

Computing

We will use the R statistical computing environment. See http://www.R-project.org. R is freely available for all common computing platforms, including Linux distributions, Mac OS X, and Windows.

Having taken a class like Stat 133 will help considerably.

Here are some good slides about R. (From Cari Kaufman's Stat 133.) The URL is

http://www.stat.berkeley.edu/~cgk/teaching/assets/Stat133AllLectures.pdf

You are free to use Python if you'd like.

Work groups

Later in the semester, you will have to form groups of three for the purpose of carrying out the final project.

Homework

A number of assignments will be due over the semester – either on a weekly or a bi-weekly basis. Working with data in R is an essential component of this course and will be part of the homework. Assignments will also check your understanding of the theory behind the methodologies we cover.

You are free to discuss the homework with other students. However, you need to write your own individual solution to the homework. Turning in a solution that is essentially identical to that of another student is not acceptable.

No extensions to due dates will be given.

Homework will be due on Friday at 3pm in the Stat office, 367 Evans.

Exams

There will be a midterm examination. I will indicate clearly which topics the exam will cover. The GSI will devote a discussion section to preparation and review for the exam. The midterm will take place the week of 3/7. There will most likely be an in-class part and a take-home, data analytic part.

It is likely that there will be a final exam: it will most likely be an oral exam – depending on enrollment at the end of class.

Final project

The final project will be a competition among the groups in the course to produce the best prediction rule on a contest dataset. Every group will write and submit a report describing exactly how it analyzed the contest data and obtained its results. The final project grade will *not* depend on your standing in the competition, but instead on the quality of the analyses attempted and of the written report. Each group member must participate in both the data analysis and the report writing; the report must include an attribution section indicating who analyzed and wrote what.

Around the beginning of April, I will give you the data.

Grading

Homework: 30% In-class exams: 40% Final project: 30% is the tentative split.

Topics

Readings from the text will be supplemented occasionally with handouts. "HTF" indicates the Hastie, Tibshirani, and Friedman text.

Here is a tentative syllabus (substantial changes are still possible)

Text

- Introduction/overview HTF1; 1 lecture
- Exploratory data analysis and unsupervised learning (pca, clustering, spectral clustering): 4 lectures;
- Supervised learning foundations: HTF2; 3 lectures
- Linear regression methods: HTF3; 3 lectures
- Linear classification methods: HTF4; 2 lectures
- Basis expansions: HTF5; 2 lectures
- Model selection HTF7; 1 lecture
- CART: HTF 9.2; 2 lectures
- Boosting and stagewise additive models: HTF10; 3 lectures
- The bootstrap: 1 lecture;
- Random Forests: HTF 15; 1 lecture
- Kernel methods (including SVMs): 3 lectures
- Neural nets and projection pursuit regression: 2 lectures

We may touch upon other topics depending on how the class goes and the interest of the students.