Statistics 154: Modern Statistical Prediction and Machine Learning

Syllabus, Spring 2014

Lectures

Time: Tue/Thu, 3:30 PM - 5:00 PM Place: 534 Davis

Lab section

Time: Friday, 9-11 and 12-2, 340 Evans

Instructor

Noureddine El Karoui

nkaroui@berkeley.edu
Office hours: Thursday, 11:00-12:00, 311 Evans

Graduate student instructor

Derek Bean

Office hours: TBA

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)

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

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

• 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

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.

No extensions to due dates will be given.

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 3/10. 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: either standard written exam or oral exams – 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
- 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
- Random Forests: HTF 15; 1 lecture
- Kernel methods (including SVMs): 4 lectures
- Unsupervised learning

Other topics we may touch upon include: robust methods (coming from robust optimization), unsupervised learning methods, and neural nets. This will depend on how the class goes and the interest of the students.