Statistics 151B
Modern Statistical Prediction and Machine Learning
Syllabus, Spring 2012

Administrative Information

Course homepage
`http://www.stat.berkeley.edu/~jon/stat-151b-spring-2012`
Announcements, handouts, problem sets, solutions, other materials.

Lectures
Time: Tue/Thu, 2:00 PM - 3:30 PM
Place: 534 Davis Hall

Discussion section
Time: Mon, 2:00 PM - 4:00 PM
Place: 170 Barrows Hall

Instructor
Jon McAuliffe 449 Evans Hall
`jon@stat.berkeley.edu`
Office hours: Thu, 12:30 PM - 1:30 PM, and by appointment

Graduate student instructor
Jeff Regier 323 Evans Hall
`jeff@stat.berkeley.edu`
Office hours: Tue 1:00 PM - 2:00 PM and Thu 3:30 PM - 5 PM, in 432 Evans

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 course website.

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

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.

(continued)

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.

Work groups
Please form groups of three for the purpose of carrying out the homework assignments and the final project. Indicate the members of your group in an email to the instructor and GSI no later than Tuesday, January 24th. If you are having trouble getting a group together, let us know *before* January 24th.

Homework
A number of assignments will be due over the semester. 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.
Each group member must participate in both solving problems and writing up solutions. In the solution write-up, briefly state who did what.
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.
There is no final exam.

Final project
We're going to try something new for the final project this year: 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 open the competition web page on Kaggle, `http://inclass.kaggle.com`. There will be a leaderboard to motivate/infuriate you. The grand prize is... respect, Don Corleone, respect. The course staff will enter the competition too. Can you beat us?

You may opt out of the competition, instead choosing a dataset to analyze and questions to investigate according to your own interests—subject to my approval of a written proposal. This option is intended for students with a substantial alternative dataset in mind. Proposals to analyze toy datasets (e.g. from the UCI repository) will be rejected.

Grading
Homework: 30%
Midterm: 30%
Final project: 40%

(continued)

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

Lec#	Date	Topic	Text
01	T 17 Jan	Introduction/overview	HTF1
02	R 19 Jan	Supervised learning foundations: I	HTF2
03	T 24 Jan	Supervised learning foundations: II	HTF2
04	R 26 Jan	Supervised learning foundations: III	HTF2
05	T 31 Jan	Linear regression methods: I	HTF3
06	R 2 Feb	Linear regression methods: II	HTF3
07	T 7 Feb	Linear regression methods: III	HTF3
08	R 9 Feb	Linear classification methods: I	HTF4
09	T 14 Feb	Linear classification methods: II	HTF4
10	R 16 Feb	Basis expansions: I	HTF5
11	T 21 Feb	Basis expansions: II	HTF5
12	R 23 Feb	Model selection	HTF7
13	T 28 Feb	Classification and regression trees: I	HTF9.2
14	R 1 Mar	Classification and regression trees: II	HTF9.2
15	T 6 Mar	Boosting and stagewise additive models: I	HTF10
16	R 8 Mar	Boosting and stagewise additive models: II	HTF10
17	T 13 Mar	Boosting and stagewise additive models: III	HTF10
18	R 15 Mar	Case study 1: face detection in real-time video	
19	T 20 Mar	Catch-up	
	R 22 Mar	*Midterm examination—in class*	
	T 27 Mar	*No lecture—Spring break*	
	R 29 Mar	*No lecture—Spring break*	
20	T 3 Apr	The bootstrap	
21	R 5 Apr	Bootstrapped prediction rules	HTF15
	R 5 Apr	*(Optional) final project proposal due—in class*	
22	T 10 Apr	Kernel methods: I	
23	R 12 Apr	Kernel methods: II	
24	T 17 Apr	Kernel methods: III	
25	R 19 Apr	Kernel methods: support vector machines	
26	T 24 Apr	Case study 2: the Netflix prize	
27	R 26 Apr	Catch-up	
	F 11 May	*Final project due—4pm, Statistics dept.*	