

Statistics 471

Modern Applied Statistics and Machine Learning

Syllabus, Spring 2007

Administrative Information

Course homepage

<http://stat.wharton.upenn.edu/~mcjon/stat-471>
Announcements, handouts, problem sets, solutions, other materials.

Lectures

Time: Tue/Thu, 01:30 PM - 03:00 PM
Place: 245 Huntsman Hall

Instructor

Jon McAuliffe 467 Huntsman Hall
mcjon@wharton.upenn.edu 215-898-8007
Office hours: Tue/Thu, 10:30 AM - 12:00 PM, and by appointment

Teaching assistant

Chaitra Nagaraja 427.4 Huntsman Hall
chaitra@wharton.upenn.edu 215-898-8222
Office hours: Tue 3:00 PM - 5:00 PM, and by appointment

Additional help

Visit the StatLab: location and schedule at
<http://stat.wharton.upenn.edu/~juntianx/statlab.html>

Required texts

Peter Dalggaard, *Introductory Statistics with R*
Trevor Hastie, Robert Tibshirani, and Jerome Friedman,
The Elements of Statistical Learning: Data Mining, Inference, and Prediction

Prerequisites

- A semester of linear algebra (matrices, vector spaces)
- A semester of multivariate calculus (in particular, partial derivatives)
- A semester of statistical inference

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 most computing platforms, including Linux distributions, Mac OS X, and Windows.

(continued)

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.

No extensions to due dates will be given.

Final project

At the end of the course you will carry out a substantial data analysis project. Your analysis will be graded based on a report you write and submit. You will choose a data set to analyze and questions to investigate according to your own interests, subject to my approval of a written proposal. The project will make use of ideas and methods presented over the semester. You may work in groups of at most three, but collaboration is by no means required. If you choose to collaborate, each person must participate in both the data analysis and the report writing; the report must then include an attribution section indicating who analyzed and wrote what.

Exams

None.

Grading

Homework: 50%

Final project: 50%

(continued)

Readings from the text will be supplemented occasionally with handouts. “D” indicates the Dalgaard text, “HTF” the Hastie, Tibshirani, and Friedman text.

Lec#	Date	Topic	Text
01	T 9 Jan	Introduction/overview	HTF1
02	R 11 Jan	Computing with data: I	D1
03	T 16 Jan	Computing with data: II	D2, D3, D4
04	R 18 Jan	Computing with data: III	D5, D6, D7
05	T 23 Jan	Computing with data: IV	D9, D10, D11
06	R 25 Jan	Supervised learning overview: I	HTF2
07	T 30 Jan	Supervised learning overview: II	HTF2
08	R 1 Feb	Linear regression methods: I	HTF3
09	T 6 Feb	Linear regression methods: II	HTF3
10	R 8 Feb	Linear regression methods: III	HTF3
11	T 13 Feb	Linear classification methods: I	HTF4
12	R 15 Feb	Linear classification methods: II	HTF4
13	T 20 Feb	Splines and basis expansions: I	HTF5
14	R 22 Feb	Splines and basis expansions: II	HTF5
15	T 27 Feb	Splines and basis expansions: III	HTF5
16	R 1 Mar	Catch-up and review	
	T 6 Mar	<i>No lecture—Spring break</i>	
	R 8 Mar	<i>No lecture—Spring break</i>	
17	T 13 Mar	The bootstrap: I	
18	R 15 Mar	The bootstrap: II	
19	T 20 Mar	Model selection	HTF7
20	R 22 Mar	Classification and regression trees	HTF9.2
21	T 27 Mar	Boosting and stagewise additive models: I	HTF10
22	R 29 Mar	Boosting and stagewise additive models: II	HTF10
	T 3 Apr	<i>No lecture—Passover</i>	
23	R 5 Apr	Boosting and stagewise additive models: III	HTF10
24	T 10 Apr	Support vector machines: I	HTF12
	T 10 Apr	<i>Final project proposal due—in class</i>	
25	R 12 Apr	Support vector machines: II	HTF12
26	T 17 Apr	Support vector machines: III	HTF12
27	R 19 Apr	Catch-up and review	
	F 04 May	<i>Final project due—3pm, Statistics dept.</i>	