

Probability and Mathematical Statistics II

Department of Statistics 36-626, Spring 2008

Lectures: MWF 10.30-11.20 AM; PH 125B

Instructor: Ann B. Lee

Baker Hall 229 J; 268-7831; annlee@cmu.edu

Office hours: Tuesdays 4–5 PM

Teaching assistant: Avranil Sarkar

Office hours: TBA; avranils@stat.cmu.edu

Main Text: Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer 2004 (Corrected second printing 2005).

Supplementary material, to be handed out in class.

Other references: Hastie et al, *The Elements of Statistical Learning*, Springer 2001.

Peter Dalgaard, *Introductory Statistics with R*, Springer 2002.

J. Maindonald and J. Braun, *Data analysis and graphics using R*, Cambridge 2003.

This course is a fast-paced and intense introduction to the statistical foundations of models and methods in modern data analysis and inference. It is primarily intended for master's level students in statistics, and PhD and master's level students in computer science and related fields. The class is a suitable preparation for advanced studies in machine learning and applied statistics.

Prerequisites are 36-625 or 36-705 in Intermediate Statistics for graduate students. The class assumes a solid background in linear algebra and multivariable calculus, including Taylor series, orthogonal expansions and eigenvalues/eigenvectors.

Course Work: There will be weekly homework assignments, one midterm, and a final exam. Homework assignments will be a mixture of theory and practical exercises; you may have to read ahead to do some of the problems. I encourage you to discuss homework problems with other students but do **not** copy other students' assignments. In other words, work together but write up your solution on your own.

Some problems will involve computing. I recommend that you use a high-level computer language such as R. Some R code and documentation will be posted on the course web site.

Grading policy: The term grade will be based 30% on the homework, 30% on the midterm and 40% on the final. Homework assignments are due on Fridays in class (if you are unable to make it to class, please slip your homework under my door before class starts). No late homework is accepted; to compensate for this stricter policy, I will discard of your lowest homework grade. Make-up exams or extensions will only be given with a note from your advisor or dean.

Tentative Course Schedule (*)

Regression, Density Estimation and Classification

- Week of 1/14 Review of fundamental concepts in inference. Course outline.
Linear regression (least squares and maximum likelihood).
- Week of 1/21 More on linear regression. Multiple regression and model selection.
- Week of 1/28 Model complexity, smoothing and the bias-variance tradeoff.
Introduction to non-parametric curve estimation.
- Week of 2/4 Histograms, kernel density estimation and non-parametric regression.
- Week of 2/11 Smoothing using orthogonal functions: density estimation, regression,
and wavelet smoothing.
- Week of 2/18 Classification: error rates and the Bayes classifier, Gaussian and linear classifiers
logistic regression, k-nearest neighbors and cross-validation.

Multivariate Models, Inference about Independence and Causality, Graphical Models.

- Week of 2/25 *Midterm exam (2/27 Wed 6-8 PM)*
Review of conditional independence and hypothesis testing.
- Week of 3/3 Multivariate models. *No class on Friday.*
- Week of 3/10 *Spring Break*
- Week of 3/17 Inference about independence.
- Week of 3/24 Causal inference.
- Week of 3/31 Directed graphs and conditional independence.
- Week of 4/7 Undirected graphs.
- Week of 4/14 Topics as time permits. *No class on Friday.*
- Week of 4/21 Topics as time permits (e.g. dynamic programming, MCMC, log-linear models)
- Week of 4/28 Review and wrap-up.
- 5/XX/06 *Final exam to be scheduled by HUB*

(*) As a rule, the lectures and the homework problems define the course.