

Statistics 640
Spring, 2009

Data Mining and Statistical Learning

Professor David W. Scott
2057 Duncan Hall, 713-348-6037, scottdw@rice.edu

Tu/Th 2:30-3:50 Keck Hall - 101
Office Hours: M 11-12 & W 1:30-2:30 or by appt.

Organizational note: *Lectures begin on Tuesday, January 6. Since many students will be taking A exams, we will discuss general material initially. The first homework assignment is given on the second page.*

Course description: Survey of ideas, methods, and tools for analyzing large data sets; techniques for searching for unexpected relationships in data. Topics from supervised and unsupervised learning include regression, discriminant analysis, kernels, model selection, bootstrapping, trees, MARS, boosting, classification, clustering, neural networks, SVM, association rules, principal curves, multidimensional scaling, and projection pursuit. (This list may be extended depending upon contents of the new edition of our textbook.)

Only an introductory background in probability and statistics is required, although some basic knowledge of classical multivariate statistical methods will be helpful. *However, no previous experience with multivariate analysis is assumed. We will cover such material in this course.* Students with particular research problems are especially welcomed. The instructor intends to accommodate students from other disciplines who are interested in this topic.

Grading: Problem homeworks; Data analysis homeworks; final project (written report and an oral presentation).

Text: Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York (second edition). The bookstore will not have the new edition until early February. (I have postscript copies of the first few chapters. Check in the course web site; see below. These were furnished as a favor by the editor, so please do not distribute and remove them when you get your book. Note that the book store is getting a shipment directly from the printing house, which will be 2 weeks earlier than anyone else will have it. I do not know what the price will be at this time.)