

Spring 2010
Statistics 702: Data Mining Statistical Methods

I. General Information

Lecture: TTh 2:00-3:15 pm, GMCS 325
Course web page: rohan.sdsu.edu/~jjfan/sta702

Instructor: Juanjuan Fan
Office: GMCS 519
E-mail: jjfan@sciences.sdsu.edu
Office Hours: T 1:00-1:50 pm, Th 11-11:50 am

Textbook: *Statistical Learning from a Regression Perspective*
by Berk (2008), Springer

References: 1. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, by Izenman (2008), Springer
2. *Data Analysis Using Regression and Multilevel/Hierarchical Models*
by Gelman and Hill (2006), Cambridge University Press

Prerequisites: Stat 551B or 670B

Grading: Assignments and group project: 20%
Midterm exams: 40%
Class participation and presentation: 15%
Final project: 25%

Note that any regrading request on homework or exams has to be made within one week after the paper is returned.

Exams and projects:

Late papers will not be accepted.
NO early or makeup exams are given - no exceptions.

II. Other Information

1. All assignments and solutions (for close-ended problems) will be posted on the course web page. There will be one group project.
2. There will be two in-class exams on Thursdays, February 25 and April 8.

3. There will be a final project (in lieu of a final exam) so that you can apply the data mining techniques to real data. You will be asked to present your findings during the last two weeks of classes. In addition, a written report will be due by Thursday, May 20 at 1-3 pm.
4. Due to concerns of computer viruses, I ask that you hand in all your assignments/reports as a hard copy. No electronic copies will be accepted.

III. Course Content

Introduction to data mining and introduction to R, cluster analysis, classification and regression trees (CART), bagging, random forests, and boosting.