Statistics 202: Statistical Aspects of Data Mining (Fall 2005)

Instructor: Jerome H. Friedman

Place / Time: Gates B1 / MW 2:45 - 4:00pm.

Description:

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Statistics, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated analysis and exploration of (usually) large complex data sets. Data Mining is having a major impact in business, industry, and science. This course covers some of the principal methods used for Data Mining, with the goal of placing them in common perspective and providing a unifying overview.

Topics:

Introduction:

What is DM? Myths: what it can and can't do. Description vs. prediction. Knowledge discovery "process".

Overview:

What is data: types of measurements. What are "patterns" in data? Statistical inference. Description vs. prediction. Types of data. Types of procedures.

Methodology:

- Decision tree induction: CART, CHAID, C4.5.
- Multiple tree models: bagging and boosting.
- Instance-based learning: near neighbor and kernel methods
- Association rules: Market basket analysis.
- Clustering: Hierarchical, K-means, mixture modeling.

Prerequisites: A familiarity with the basic concepts in probability, calculus, linear algebra, and optimization. Statistics116 useful (not required).

Logistics:

- Office hour: After class and/or by appointment.
- TAs:
  - Sarah Emerson, sarah.emerson@stanford.edu, Office Hours: Thursday 3:30pm - 5:30pm, Sequoia Hall 223A.
  - Yaqian Guo, yaqiang@stanford.edu, Office Hours: Wednesday 4-6pm.
  - Gen Nowak, gnowak@stanford.edu, Office Hours: Tuesday 2pm - 4pm, Sequoia Hall 223A.
  - Victoria Stodden, vcs@stanford.edu, Office Hours: Friday 11am - 1pm, Sequoia Hall 227.
• Homework:
  ○ reading assignments (with questions).
  ○ computing assignments (apply / implement methods - discuss results).
• Midterm:
  ○ maybe -doubt it.
• Final:
  ○ likely - not sure what form.