CS395T/CAM395T: Large-Scale Data Mining
Fall 2003

Class times: MW: 4-5:30 pm, GEO 2.102
Instructor: Inderjit Dhillon
Homepage: http://www.cs.utexas.edu/users/inderjit
E-mail: inderjit@cs.utexas.edu
Office: ACES 2.332, Ph: 471-9725
Office Hours: Wed 2-3pm. Other times by appointment.
Class URL: http://www.cs.utexas.edu/users/inderjit/courses/dm2003/
TA: To be announced

Prerequisites: Basics (undergraduate level) of linear algebra and statistics, or consent of the instructor.

Outline: Recent times have seen an explosive growth in the raw data available electronically. The data occurs in various forms, e.g., as text, image, video or numeric data; both in public domains and in private corporations. Data mining is the automatic discovery of interesting patterns and relationships in very large data sets. This graduate course will focus on scalable algorithms for data mining, with an emphasis on applications in text mining and bioinformatics. Topics covered will include algorithms for (a) *search* using link analysis (Google, HITS), (b) *clustering*, (c) *classification*, (d) *factor analysis*, and (e) *multidimensional visualization*. The technical tools used in the course will come from: linear algebra, multivariate statistics, information theory and optimization. The concepts used from these technical areas will be covered in class, but undergraduate level linear algebra is a pre-requisite (see above).

A substantial portion of the course will concentrate on research projects, where students will choose a well-defined problem. Students will have freedom in choosing projects, as long they are related to data mining. Projects can vary in their theoretical/mathematical content, and in the amount of implementation/programming involved. Projects will be conducted by teams of up to 2-3 students.

Grading:
10 + 30 pts: Class Project (First submission + Final submission).
20 pts: Four Homeworks.
25 pts: One midterm exam.
10 pts: Class presentation of a research paper.
5 pts: Attendance and participation in class discussions.

References:
Elements of Statistical Learning: Data Mining, Inference, and Prediction by T. Hastie, R. Tibshirani, J. Friedman, Springer-Verlag, 2001.
Pattern Classification by R. Duda, P. Hart and D. Stork, John Wiley and Sons, November 2000.
Foundations of Statistical Natural Language Processing by C. Manning and H. Schutze, MIT Press, 1999.

Disabilities statement: "The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TTY."