CS 395T: Large-Scale Data Mining
Unique No. 49460
Spring 2000

Class times: MW: 4:00 - 5:30 pm, Welch Hall 3.260
Instructor: Inderjit Dhillon.
E-mail: inderjit@cs.utexas.edu
Office: Taylor Hall 5.148, Ph: 471-9725
Office Hrs: Wed 10:00-11:00am. Other times by appointment *only*.
Class Homepage: http://www.cs.utexas.edu/users/inderjit/courses/datamining.html

TA: Shailesh Kumar.
E-mail: skumar@cs.utexas.edu
Office: ENS 518
Office Hours: Th 10-1pm.

PREREQUISITES: Graduate standing OR consent of the instructor. You are expected to know basics (undergraduate level) of linear algebra and some statistics.

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. A special emphasis will be on information retrieval, especially for the World Wide Web. Topics covered will include (i) link analysis on the internet (like Google), (ii) content analysis of documents using SVD (Singular Value Decomposition), SVMs (Support Vector Machines), and linear discriminant analysis, (iii) image segmentation using graph partitioning, (iv) face detection in images, (v) clustering and classification algorithms, (vi) visualization of high-dimensional data, etc. We may study other application areas, such as bioinformatics, if there is sufficient interest and need.

A substantial portion of this class will be paper readings and research projects, where students will have freedom of choosing a well-defined problem of their choice. Emphasis will be on implementing algorithms that could lay the foundations of a usable system, such as, a text categorizer. An elementary knowledge of linear algebra would be helpful but is not essential.

FORMAT: Lectures + Paper Presentations.

GRADING:
25 + 40 pts: Class Projects (First submission just after Spring Break).
20 pts: Paper Presentation
15 pts: Class Participation.
There will be no midterm or final exam.

TEXTBOOKS:
Pattern Classification and Scene Analysis by R. Duda and P. Hart, Wiley-Interscience, 1973.
Foundations of Statistical Natural Language Processing by C. Manning and H. Schutze, MIT Press, 1999.
Buying these books is not essential. They will be placed on reserve at PCL.