Guidelines and Deadlines for Data Mining Term Project

1. Guidelines

You are encouraged to work in groups of two to four for the term project. This project is a critical part of the course, and a significant factor in determining your grade. First, each group should submit a one/two page proposal summarizing the proposed project including the plan of attack and at least two key references, on or before March 26th, so that I can comment if needed. Please feel free to discuss your project with me before that. Also, if you have difficulty in finding a project partner, you could contact me with a list of your interest area(s), so I can try to match you up with another student during our March meeting. Each group shall give a 15–20 minute presentation on their project on 14/15th May.

The **deadline** for the hardcopy term paper submission is **Friday, May 14th.** This copy should be 10-30 pages (1.5 spacing) including figures, tables and/or references. If a part of this is an html file/directory, you could just give me a pointer to the URL, and reduce your printed submission accordingly. One submission per group. In addition I'd like a softcopy (.ps/.pdf/html/word) provided as a single file or link, mailed to the TA, to be kept at the secure course site for posterity. I'll try to provide feedback on the final term paper the next day, or by email soon thereafter.

2. Project Types

The project can be a practical one (a specific application of data mining), or a theoretical one (propose a new algorithm etc). In-depth survey papers on new or focussed topics are also possible, but it should be an original and unique niche (several broad surveys exist on the web; a cut-and-paste job will not do!). You can choose any topic relating to data mining, and use any reasonably large data set. Some tools and datasets are referred to at the course website; also see kdnuggets.com. If you have access to an interesting data set at work (and equally important, have (or have access to) domain knowledge about the data, then you can analyze such a set for your project. I can sign NDA if need be.

Below are some suggested topics; by using Google and **Citeseer** http://inquirus.nj.nec.com/cs, you should be able to find some materials on them.

Clustering: - clustering with constraints; clustering very high dimensional data; clustering of gene sequences; etc, see http://lans.ece.utexas.edu/ ghosh/ch8s.ps for a review and other issues.
- clustering tools: (a) CLUTO (Karypis): understand this tool; add to its functionality, or
(b) Co-clustering: understand this approach; add to its functionality.
http://www.cs.utexas.edu/users/yguan/datamining/cocluster.html

Classification: - comparing different approaches to solving multi-class problems: error correcting output codes (Bakiri & Dietterich) vs. Binary Hierarchical Classifier (Kumar, Ghosh, Crawford).
- bagging with strong learners vs. boosting with weak learners
- dealing with highly different costs; priors (Charles Elkan, Domingos)

Bioinformatics: Hot topic! e.g. see the KDD2002 challenge competition
http://www.biostat.wisc.edu/ craven/kddcup/ ; also my summer project offering (1.C).

Privacy Perserving Data Mining: e.g. see paper No.2 at
http://www.lans.ece.utexas.edu/ srujana/papers.html, and the special issue of SIGKDD:
http://www.acm.org/sigkdd/explorations/issue4-2.htm

Or you can consider Distributed Data Mining: How to do data mining (regression, classification, AR, clustering) over multiple, geographically distributed and possibly quite varying, data sets.

Remote Sensing: (i) predicting forest cover type from satellite images (see UCI).
(ii) classifying land cover from hyper-spectral data (I can provide).
(iii) discovering and modelling heterogenous regions in spatial data

(e.g. http://www.cs.umn.edu/research/shashi-group/ ;
http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/SpatialKDD/) more generally modeling signals with distinct regimes (determine boundaries; model each segment).

Web Mining: There is a fairly extensive list of projects provided on web mining (along with references, etc) at http://lans.ece.utexas.edu/course/prac/03sp/prac-projects.html Also the KDD Cup for 2003 related to citation prediction, see http://www.cs.cornell.edu/projects/kddcup/

Streaming Data; Change Detection Analyzing data that you see only once (Johannes Gehrke); Detecting and modeling "change" in time sequence data, e.g. those gathered from networked computer systems; change in customer segmentation (how do clusters move, get created/die, as the stats of data gathered changes over time?).

Direct Marketing: donor database (KDD cup 1998), from UCI

Statistical Angles Explore role of sampling techniques; make connections between statistical and machine learning approaches, e.g. see Trevor Hastie's works.

Semiconductor Manufacturing: measuring quality of chips, equipment, process etc. from process logs. (you need to get your own data).

Intrusion Detection: several papers recently.. build a network intrusion detector, a predictive model capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections. Detecting and modeling "change" in networked computer systems.