EE 380L-10/BME383J-7: DATA MINING

Spring 2011

Class times: TTh: 11am-12:30pm, ENS 116, Unique No. 16960(ECE)/14420(BME)
Instructor: Joydeep Ghosh, ghosh@ece.utexas.edu; www.ideal.ece.utexas.edu/~ghosh
Office: ACES 3.118, 471-8980
Office Hrs: TTh 1:30-2:30pm. Other times by appointment only.
TA/grader info: Priyank Patel; priyank.patel@mail.utexas.edu Office ENS 110, hrs: MWF 11am-noon; TTH 9:30-11am
PREREQUISITES: (Graduate standing in Engineering, CS, Maths or Physics) OR (consent of the instructor). You are expected to know basics (undergraduate level) of probability/statistics. Knowledge of basic linear algebra will help as well.
SUPPLEMENTARY COURSE URL: http://www.ideal.ece.utexas.edu/courses/ee380l/
Note, reading lists, scores etc, will be communicated via Blackboard.

COURSE OUTLINE: The information explosion of the past few years has us drowning in data but often starved of knowledge. Many companies that gather huge amounts of electronic data have now begun applying data mining techniques to their data warehouses to discover and extract pieces of information useful for making smart business decisions. Effective data mining, as opposed to data dredging, requires an understanding of concepts from exploratory data analysis, pattern recognition, machine learning, heterogeneous data bases, parallel processing and data visualization, in addition to knowing the problem domain.

Given the rich set of topics in this area, I’ll be concentrating on only some core topics. The tentative schedule of classes is linked here

The course is mostly a set of lectures by me, setting up basic concepts. 
There will be a couple of guest lectures by visitors from industry and academia. The last 4-5 classes will consist of student term-project presentations, followed by active discussion.

I encourage you to do a term project on one of the four focus topics: (i) Topic Models (ii) Predictive modeling of multi-relational data, (iii) Distributed Data Mining/Learning and (iv) Data Mining for Health Care (involving a large data set).
If you want to work on a topic outside of this list, please check with me first. At the very least, it should involve a large enough data set (e.g. “p” at least 20, and “n” times “p” over a million).
I will need a preproposal from each group by March 10.

GRADING:
10+30 pts: Project. (groups of 2-3). (15-25 min. presentation + Term paper due May 5)
25 pts: homeworks, including paper/topic critiques.
5 pts: Pop-quiz (late Feb).
25 pts: Written Exam; Tues, March 29, in class
5 pts: Participation in discussions.

There will be no final exam.
A set of class notes and supplementary materials will be available via Blackboard.

**Textbook**

In addition, my notes will be available via Blackboard, and a reading list of papers will also be provided.

**Some other recommended books:**

**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."
The above was a mandated statement, quoted verbatim. It does not imply that this course is disabled. I wonder what TTY means.

**WEBSITES:**
- [Data Mining Web Sites](http://hercules.ece.utexas.edu/~ghosh/dm11.html)
- [Data Mining and Knowledge Discovery Resources](http://hercules.ece.utexas.edu/~ghosh/dm11.html)