MOTIVATION AND GOALS:

We live in the age of information and knowledge management. The importance of collecting data that reflects business or scientific activities to achieve competitive advantage is widely recognized today. Advanced systems for collecting data and managing it in large databases are in place in most large and mid-range companies. However, the bottleneck of turning this data into your success is the difficulty of extracting knowledge about the system from the collected data.

Below are some of the questions that can be answered if information hidden in a database can be found explicitly and utilized:

• What goods should be promoted to this customer?
• What is the probability that a certain customer will respond to a planned promotion?
• Can one predict the most profitable securities to buy/sell during the next trading session?
• Will this customer default on a loan or pay back on schedule?
• What medical diagnosis should be assigned to this patient?
• How large are the peak loads of a telephone or energy network going to be?
• Why does the manufacturing facility suddenly starts to produce defective goods?

Modeling the investigated system and discovering relations that connect variables are the subject of data mining.

Modern computer data mining systems self learn from the previous history of the investigated system, formulating and testing hypotheses about the rules which this system obeys. When concise and valuable knowledge about the system of interest is discovered,
it can and should be incorporated into some decision support system which helps the manager to make wise and informed business decisions.

COURSE OVERVIEW

The course will introduce concepts and techniques of data mining and data warehousing, including concepts, principles, architectures, designs, implementations, and applications of data warehousing and data mining.

TOPICS:

- Introduction
- Data warehousing and OLAP technology for data mining
- Data preprocessing
- Descriptive data mining: characterization and comparison
- Association analysis
- Classification and prediction
- Cluster analysis
- Mining complex types of data
- Applications and trends in data mining

MECHANICS

You must be enrolled to attend the lectures.

TEXTBOOK(S)

(1) Data Mining: Concepts and Techniques
    Jiawei Han, Micheline Kamber

(2) Microsoft SQL Server 2008 Analysis Services Step by Step
    Microsoft Press; 1st Edition (4/09)

PREREQUISITES

Students enrolling in this class should have taken introductory courses in databases and fundamental algorithms. Knowledge or experience in data warehousing, working knowledge of a mainstream
database system (e.g., Microsoft SQL Server, IBM DB2, Oracle 11g), and previous programming experience in at least one higher-level procedural or object-oriented language are a plus.

REFERENCES

- Investigative Data Mining for Security and Criminal Detection, Jesus Mena, Butterworth-Heinemann, 2003
- Business Modeling and Data Mining, Dorian Pyle, Morgan Kaufmann, 2003
- Predictive Data Mining by S.M. Weiss and N. Indurkhya
- Seven Methods for Transforming Corporate Data Into Business Intelligence by Vasant Dhar
- Data Mining Techniques: For Marketing, Sales, and Customer Support by Michael Berry
- Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations by Ian Witten; Morgan Kaufmann, 1999
- The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses by Ralph Kimball
- Data Warehouse Project Management by Sid Adelman
- Data Warehouse: From Architecture to Implementation by Barry Devlin

SOFTWARE

Microsoft SQL Server 2008 Analysis Services

Oracle Data Mining

A freely available data mining software such as IlliMine (http://illimine.cs.uiuc.edu/), DB Miner (http://www.dbminer.com/), DB2 Intelligent Miner (www.ibm.com/university), etc.

Weka Data Analysis and Mining software
(http://www.cs.waikato.ac.nz/-ml/weka/)

R system for statistics with data mining algorithms
(http://www.r-project.org/)
Additional software references for further experiments and/or class project:

- IBM OLAP Miner
- Teradata Warehouse Miner
- ESRI Spatial Business Analyst
- MS Analytics
- SAS Enterprise Miner

REQUIREMENTS

Three homeworks (25%), Class participation (10%), Projects (35%), Final (30%) (Tentative). The project report is due at the last day of the class. There will be no final exam.

Collaboration on the problem sets is allowed. You may work together with one or two other partners and sign your names to a single submitted homework. All team members will receive the grade that the homework merits. There is no penalty for working on problem sets in teams of up to three (more than three is not allowed).

COURSE PROJECT

The course project is an opportunity for student groups to investigate a data mining problem that interests them. The course project should apply data mining techniques to real-world problems. Data and software for these projects can be obtained from various Internet sites, or developed by students.

A presentation of each project is required in addition to a written report.

Sample project ideas include but are not restricted to the following:

1. Compare approaches to a particular problem on criteria as accuracy, memory utilization and performance. Implement several alternative approaches and rigorously compare them on data sets with distinct properties. You can also create artificial databases to test the bounds of each approach. Some comparisons include comparing characterization methods, feature selection methods, clustering methods, and parallel data mining approaches.
2. If you're interested in working as a data analyst, you are encouraged to study real world problems and needs for data mining. Use whatever means you can find to discover interesting patterns. The following are suggested topics: Customer segmentation, Predictive model for customer retention, Customer churn in Teleco, Mining of Web logs and fraud detection.

An example is data from the 1998 KDD Cup data mining contest (http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html). The overall task is to develop a predictive model that selects optimally which individuals should be sent a donation request. The details of the task are described in the contest instructions.

You should implement and test at least two different methods for solving this problem. You do not need to use complex classification and regression algorithms for this task. A combination of naive Bayesian learning, linear regression, and bagging would be fine, for example. You may implement your own software or use or modify software that you obtain elsewhere (recommended).

Another example is predicting prices of initial public offerings. Determine how much you should pay for an initial public offering on the first day of offering. Check (http://www.cs.utsa.edu/~kwek/cs4793/ipoDescription.txt) for further description.

A final example is weather data. There is an abundance of weather data online. In particular, the National Climatic Data Center (http://www.ncdc.noaa.gov/) has some free datasets online. One of the data mining tasks you might try is to predict the weather at a given time period from previous time periods. Another task you might try is clustering to partition a region into different climates.

3. A survey/research paper that can lead to a tutorial material or a journal paper. Extensiveness, comprehensibility, technical worthiness are major considerations. You should only choose this type of projects if you are familiar with the subfield you wish to survey; otherwise, you're advised not to do it. The following list of topics is suggested for your reference. Topics may include: Web usage mining, Data mining and E-business, Mining unstructured and semi-structured data on WWW, Text mining, Spatial data mining,
Multimedia data mining, Content-Based Image Indexing and Retrieval, and Data mining applications in finance.

WEB SITES

Data Mining Journals
KDnuggets
Decision Support Systems
Data Mine
UCI Data Sets
DBWorld
ACM SIGKDD
ACM SIGMOD
Data Mining and Knowledge Discovery Journal
IBM Academic Initiative
Oracle Technology Network

OTHER RECOMMENDATIONS

Students are encouraged to review the references provided on the course Web site, subscribe to Application Development Trends (www.adtmag.com), Intelligent Enterprise (http://www.intelligententerprise.com/), Information Management (http://www.information-management.com/), and Dr.Dobb's (http://www.ddj.com/web-development/).

SCHEDULE (Tentative - Guest Talks Sessions not included)

<table>
<thead>
<tr>
<th>Week 1 – Session 1</th>
<th>Course Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 2 – Session 2</td>
<td>Data Mining Introduction</td>
</tr>
<tr>
<td></td>
<td>Assignment 0 due</td>
</tr>
<tr>
<td>Week 3 – Session 3</td>
<td>Data Preprocessing</td>
</tr>
<tr>
<td>Week 4 – Session 4</td>
<td>Data Warehousing and OLAP</td>
</tr>
<tr>
<td></td>
<td>Assignment 1 due (Textbook Questions)</td>
</tr>
<tr>
<td>Week 5 – Session 5</td>
<td>Characterization</td>
</tr>
<tr>
<td>Week 6- Session 6</td>
<td>Association</td>
</tr>
<tr>
<td></td>
<td>Assignment 2 due (Data Warehousing Practice)</td>
</tr>
<tr>
<td></td>
<td>Group Project proposal due</td>
</tr>
<tr>
<td>Week 7 – Session 7</td>
<td>Classification</td>
</tr>
</tbody>
</table>
| Week 8 – Session 8 | Classification (cont'd)  
*Group Project design due* |
| Week 9 – Session 9 | Clustering |
| Week 10 – Session 10 | Data Mining Applications (e.g., Text Mining, Data Streams, and Time Series)  
*Assignment 3 due (Weka Practice)* |
| Week 11 – Session 11 | Student Group Presentations  
*Part 1 / 2* |
| Week 12 – Session 12 | Student Group Presentations  
*Part 2 / 2* |

**READINGS**

Assigned readings for the course will be from the textbooks, various Software Engineering-related Web sites, trade magazines, and recommended books listed on the course Web site.

**ASSIGNMENTS**

Homework and project assignments completion will be required. Quizzes will be administered. The final exam will be a take-home exam.

**GRADING POLICY**

25% Assignments  
35% Projects  
30% Final Exam  
10% Attendance and Participation  
Extra credit will be granted periodically for particularly clever or creative solutions.