

## CSE634 Data Mining, Spring 2005 Professor Anita Wasilewska

**Meets** Monday Wednesday 2:20 pm - 3:40 pm

**Place** CS Building 1441

### **Professor Anita Wasilewska**

e-mail address: anita@cs.sunysb.edu

Office phone number: 632 8458

Office location: Computer Science Department building, office 1428.

**Professor Office Hours** Monday, Wednesday 12:30 - 2:00 pm and by appointments.

### **Textbook**

DATA MINING Concepts and Techniques

Jiawei Han, Micheline Kamber

Morgan Kaufman Publishers, 2002

**Course Description** Data Mining, called also Knowledge Discovery in Databases (KDD) is a new multidisciplinary field, It brings together research and ideas from database technology, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, information retrieval, high-performance computing, and data visualization. Its main focus is the automated extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

The course will closely follow the book and is designed to give a broad, yet in-depth overview of the Data Mining field and examine the most recognized techniques in a more rigorous detail. It also will explore the newest trends and developments of the field in form of talks based on newest research papers from the field.

**Grading** There will be:

1. **Two** in class presentation (50pts each) given individually. Students will be graded for the presentation skills, the content, organization, clarity, and amount of work put into research and preparation.
2. One Presentation report (25 pts)
3. One students presentations evaluation report (25 pts)
4. Final Paper (50pts)

**Final grade computation** During the semester you can earn 200pts or more (in the case of extra points). The grade will be determine in the following way:  $\#$  of earned points divided by 2 = % grade.

The grade will be determine in the following way: of earned points = % grade. The % grade which is translated into letter grade in a standard way i.e. 100 - 90 % is A range, 89 - 80 % is B range, 79 - 70 % is C range, 69 - 60 % is D range and F is below 60%.

**Presentations** This is an opportunity for students to learn how to deliver

1. a well structured and prepared lecture based on the textbook material,
2. understand and present a newest research paper (or data mining application).

Students will have to prepare power point based lecture slides and explain in detail, with examples the material.

The hard copy (black and white in slide spread format) of the slides and the CD containing the presentation is to be delivered to the Professor before the presentation starts.

**Presentation 1** will be based on, or extending the content of the book. Sometimes you would need some related materials to be found from other sources.

This is an **educational part**. The main goal of this presentation is to teach others the material.

Students have to put time and effort into **understanding the material**, present it slowly and be prepared to answer students questions.

Remember that "I don't understand" is also an answer, but don't over-use it! The better answer is: "the book is not very clear, I think that its is ...".

I will distribute subjects very shortly.

**Presentation 2** will be given at the end of the semester. It is a is a presentation of a research paper or Data Mining major application. The goal of this part is to bring update to what is being done on the subject today.

Students have a total freedom of choice of then subjects. This presentation must consists of two parts.

**Part 1** Short overview of a techniques, methods used as taught during the course or found in the literature.

**Part 2** Detailed presentation of then paper or application.

### **General Principles of Presentations**

**First slide** must contain your names, student ID, professor name, course number and the title.

**Second slide** must contain ALL sources you used for the LECTURE part of your presentation. The book is included. In the case of the book the reference you have to put are title of the chapter, sections and pages numbers.

**Third slide** is an OVERVIEW of your presentation.

**Fourth slide** include the title and bibliography of your sources.

**Please**, e-mail a text file containing information included on these 4 slides to me.

**Remember** to include give a source of any picture, of slides copied from a source or any DIRECT citation on the bottom of each of your slides where it appears.

**Students Presentations report** Classroom attendance is essential to the understanding of other students presentations and learning the material of the course. You are graduate students, so I will not insult you by taking the attendance. But I want everybody to submit a written report about 10 of other students presentations. The report must contain:

1. a short motivation WHY you chose those presentations for the report,
2. One page description-summary (own words!) of each presentation,
3. your own evaluation of the presentation.

**Final Paper** Here is the procedure:

**Step 1** FIND (Web or other sources) a research paper on DATA MINING subject of your choice.

**Step 2** Write motivation why you have chosen this particular paper.

**Step 3** Write at least one page summary of the paper. In your own words. Do not copy abstracts or summaries. You have to state if it is an application or theoretical paper and what is the real point of the paper. It has to be your own summary, not the author's. You have to specify which techniques, algorithms, are used or improved upon etc...

**Step 4** Write your own evaluation of the paper. Address the following:

1. Does the author(s) really accomplished what they said they did.
2. How important is the result - based on what you KNOW (after our course!) about the field.
3. How well the paper is written: motivation, description of related research, statement of the problem of the paper, its history and relevance to the field.
4. How important is the paper with respect of future development of the field: does it open new directions, or in a case of general model building paper, how much of the past research the does it cover.
5. Any other remarks and your own reflections.

**GENERAL PRINCIPLE** Any direct citations (even of ONE SENTENCE!) must have a standard form of a citation: give the page of the paper and show clearly when it start and when it finishes.

## Course Contents and Schedule

The course will follow the book very closely and in particular we will cover the following chapters and subjects. The order does not need to be sequential.

**Chapter 1** Introduction. General overview: what is Data Mining, which data, what kinds of patterns can be mined.

**Chapter 2** Data Warehouse and OLAP technology for Data Mining. (Students presentations)

**Chapter 3** Data preprocessing: data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation.

**Chapter 4** Data Mining Primitives, Languages and System Architectures. (Students presentations)

**Chapter 5** Concept Descriptions: Characteristic and Discriminant rules. Data Generalization. EXTRA: Example of decision tables and Rough Sets.

**Chapter 6** Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm.

**Chapter 7** Classification and prediction. Decision Tree Induction ID3, C4.5). Rough Sets.  
Bayesian Classification. (Students presentations).  
Classification based on Concepts from Association rule mining.  
Genetic algorithms. (Students presentations) Statistical Prediction.

**Chapter 8** Cluster Analysis. A Categorization of major Clustering methods. Some students presentations.

**TRENDS and Developments** - newest research and applications presentation.