

CSE634, CSE590 Data Mining, Spring 2009
Professor Anita Wasilewska

Meets Tuesday Thursday 3:50 pm - 5:10 pm

Place SB Union 237

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 Tuesday, Thursday 2:00 pm - 3:13 pm and by appointments.

TA t.b.a

e-mail

TA Office hours t.b.a

Textbook

DATA MINING Concepts and Techniques

Jiawei Han, Micheline Kamber

Morgan Kaufman Publishers, 2003

Second Edition

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. Presentations 1, 2 (100pts total) given in teams of 2-4 students. The group will be graded for the presentation skills, the content, organization, clarity, and amount of work put into research and preparation. Each member of the team has to present its own well defined part and will be graded individually on this part as an overall evaluation of the group.

Presentation 1 (70pts) is a lecture type one hour presentation (see description below) given in 2-4 students groups. All members of the group must present the material in more or less equal manner.

Presentation 2 (30pts) is a short, 10-20 minutes presentation of a research paper, or an application.(see description below) given by the same group as Presentation 1. All members of the group must present the material in more or less equal manner.

Presentation 1 and 2 can be combined in one, whole class period long presentation, or can be delivered separately.

item[2. Midterm] (100pts) test covering the material from chapters 1, 2, 5, 6 included in Professor Lectures. It will be given after we finish my lectures. I plan it for the week of March 16, but it could be changed.

3. Project and Project Presentation (70pts).

4. Presentations evaluation reports (30 points).

Final grade computation During the semester you can earn 300pts or more (in the case of extra points). The grade will be determine in the following way: $\#$ of earned points divided by 3 = % 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 Principles

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

Second slide must contain ALL sources you used for the 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.

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.

HARD COPY of the presentation (black and white in slide spread format) of the slides and the CD containing the presentation is to be delivered o the Professor **before the presentation** starts. All materials must be put in a Presentation Folder labeled with students names, ID and Presentation Group number.

You receive 0-25pts for the organization of your submitted materials for presentation 1, 0-10pts for presentation 2 .

PRESENTATION Power Point FILE has to be send to course TA within 3 days of the presentation. The 3 days may be needed to do some improvements after the presentation.

ALL PRESENTATIONS will be available on the web for other students to help them to write their presentations reports. Of course students MUST attend the presentations to LEARN the material and evaluate the presentation delivery, but by having access to already delivered (and improved, if needed) presentations they will be able to to comprehend better the material and hence to judge better other students work.

Project Presentation (60pts) is a 10 minutes presentation of a **class project** that you are supposed to conduct in installments during the whole semester. See project description handout.

60pts are assigned to the quality of the project, project description and documentation and 15pts to the quality of presentation.

Presentations evaluation reports (40pts) Each student has to evaluate 10 presentations (4pts each) and submit the evaluation report. Students evaluation reports are to be **FIRST** written during presentations and submitted to Professor at the end of the class, then can be improved - and resubmitted after the presentations that were evaluated are published on the web page.

Presentation 1 main goal is to **teach others** the material. It is a detailed, lecture type presentation. It has to be based on, or extending the content of the book, book slides (if you need them come and copy from me), my slides, past students presentations slides, or other sources. Sometimes you would need to use some related materials to be found from other sources. Examine students presentations from previous course for guidance - and sources. You must **IMPROVE** on them and add your own 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 it is ..., or I understood it as ...".

Presentation 2 is a presentation of a new research, research paper, or a commercial application connected with your presentation 1.

The structure of the presentation 2 is as follows:

1. If you present a paper you must include on your first slides authors names, title and place (journal, conference) where it was published and the date of the publication, or any other source of the paper you use. **You must MAKE** a copy of the paper and put it in your **PRESENTATION FOLDER**.

If you present a commercial application you must find relevant data about the application and include it in your Presentation Folder.

General Remark You work, and will be judged as a **GROUP**.

Each group member must present some part of the whole group work. The format of how you decide to do it is left to you as a group.

PROJECT PRESENTATION This is also a group project and presentation; you work in the same group as for Presentations 1, 2. Project presentations are very short (max 10 minutes)(and will be scheduled during the last 3 classes.

Students Presentations reports Classroom attendance is essential to the understanding of other students presentations and learning the material of the course. Students take notes for their report in class, give me the preliminary sheet at the end of the class and write full report when the presentation under evaluation is on the course web page. You must submit your report to **PROFESSOR** within a week of the presentation. Each report must include:

1. one page description-summary (own words!) of the presentation content,
2. your own evaluation of the presentation.

Evaluation forms are on the course web page.

POSSIBLE PRESENTATION 1 SUBJECTS are:

Data Warehouse and OLAP technology for Data Mining.

Data Mining Primitives, Languages and System Architectures.

CRISP standards for Data Mining

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

Classification based on Concepts from Association rule mining.

Classification Accuracy testing methods and problems.

Statistical Methods 1: Statistical Prediction, Prediction by Regression, other purely statistical methods

Statistical Methods 2: Classification by Neural Networks

Statistical Methods 3: Bayesian Classification.

Statistical Methods 4: Cluster Analysis. A Categorization of major Clustering methods.

Evolutionary Computing: Genetic algorithms as optimization, Genetic algorithms as classification. Other evolutionary computing methods.

NEW ADVANCES in Data Mining, for example:

Web Mining: an overview of methods and problems

Text Mining: an overview of methods and problems

Visualization and Data Mining techniques

Natural Language Processing and Data Mining techniques

FIND YOUR OWN subject and discuss it with the Professor.

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 preprocessing: data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation.

Chapters 3, 4 Data Warehouse and OLAP technology for Data Mining. (Students presentations)

Chapter 5 Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm (LECTURE and Students Presentation).

Chapter 6 Classification and prediction.

Decision Tree Induction ID3, C4.5). Rough Sets. (Lecture and Students Presentations)

Neural Networks (Lecture and students Presentations) Bayesian Classification. (Lecture and Students presentations).

Classification based on Concepts from Association rule mining. (Students presentations).

Genetic algorithms. (Students presentations) Statistical Prediction (Students presentations).

Chapter 7 Cluster Analysis. A Categorization of major Clustering methods. (Lecture and Students presentations).

Applications and TRENDS in DM - chapters 8 -11, reading and /or students presentations.