CSE634, CSE590  Data Mining, Spring 2014
Professor Anita Wasilewska

web page: http://www.cs.stonybrook.edu/~cse634/

Meets  Tuesday Thursday  11:30 am - 12:50 pm
Place   Physics P116

Professor Anita Wasilewska
   e-mail address: anita@cs.stonybrook.edu
   Office phone number:  632 8458
   Office location: Computer Science Department building, office 1428.

Professor Office Hours  Tuesday, Thursday  2:30 pm - 3:30 pm and by appointments.

TA  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) and now called also BIG DATA 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 Components

During the semester students are responsible for the following
1. **TWO Presentations**  (100pts total)

Presentations can be given individually or in teams of 2 students.

Presenters will be graded for the presentation skills, the content, organization, clarity, and amount of work put into research and preparation form and delivery.

In case of a group presentation each member of the team has to present his/hers own well defined part and will be graded individually on this part as well as a part of overall evaluation of the group.

**Presentation 1**  (70pts)

It is a lecture type one hour long presentation (see description below)

In case of a group presentation each member of the team has to present material in more or less equal manner.

**Presentation 2**  (30pts)

It is a short, 10-20 minutes presentation of a research paper, or an application (see description below)

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

2. **Midterm/Final**  (100pts)

It is test covering the material from chapters 1, 2, 5, 6, 7 included in Professor Lectures. It will be given after we finish my lectures. I plan it for the week after the break but it could be changed.

3. **Presentations evaluation reports**  (30 points).

4. **Final Paper OR Final Project**  (70pts)

It is due anytime before Finals Week

See Project Description and email Professor when you decide which option you choose.

**Final grade computation**

Attention:

NONE of the grades will be curved!

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 which is translated into letter grade as follows.

100 - 90 % is A range; A (100-96%) A- (95- 90%), 89 - 80 % is B range: B- (80 - 82%), B (83 -85%), B+ (86 -89%), 79 - 70 % is C range: C- (70- 72%), C (73-75%), C+(76-79%), 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 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 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, and 0-10pts for presentation 2.

Presentation Power Point file (or other) 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.

Students Presentations will be available on the course webpage for other students to help them to write their final presentations reports.

Of course students should attend the presentations to learn the material and evaluate the presentation delivery.

I will collect their preliminary reports written in class during the presentations.

By having access to already delivered (and improved, if needed) presentations students will be able to comprehend better the material and hence to judge better other students work and write their final presentation report.

The final presentation report is invalid without submitted in class preliminary report.
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, or any other sources.

Presenters 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 PRINT a copy of the paper and put it in your Presentation Folder.

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

3. If you do it 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.

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.

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.

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. 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 you 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.

SOME POSSIBLE PRESENTATION 1 SUBJECTS are:

Data Warehouse and OLAP technology for Data Mining. (Chapter 3 of the Book)

Data Cube Computation and Data Generalization (Chapter 4 of the Book)

Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm. (Chapter 4 of the Book)

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 DM techniques

Natural Language Processing and DM 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.

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

2. Neural Networks (Lecture and students Presentations)

3. Bayesian Classification. (Lecture and Students presentations).

4. Classification based on Concepts from Association rule mining. (Lecture)

5. Genetic algorithms. (Lecture and Students presentations)

6. 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.

Required Syllabi Statements: The University Senate has authorized that the following required statements appear in all teaching syllabi on the Stony Brook Campus.
Americans with Disabilities Act: If you have a physical, psychological, medical or learning disability that may impact your course work, please contact Disability Support Services, ECC(Educational Communications Center) Building, Room 128, (631)632-6748. They will determine with you what accommodations, if any, are necessary and appropriate. All information and documentation is confidential.

Academic Integrity: Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person’s work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology & Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures.

CSE590/634 DATA MINING
PROJECT DESCRIPTION

BAKARY DATA - on the course web page.

This is a classification data with TYPE DE ROCHE (Rock Type) as a CLASS attribute. There are 98 records with 48 attributes and 6 classes.

Classes are:

C1 : R. Carbonates AND R. Carbonatees impures  
C2 : Pyrate  
C3 : Charcopyrite  
C4 : Galene  
C5 : Spahlerite  
C6 : Sediments terrigene

Most important attributes (as determined by the expert) are: S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3

This is a real life experimental data and it contains a lot of missing data (no value).

THE PROJECT GOAL is to use an Internet based CLASSIFICATION TOOL to generate sets of DISCRIMINANT RULES describing the content of the data. You can choose one you like, or use WEKA: http://www.cs.waikato.ac.nz/~ml/weka/index.html

The project has to follow all steps of Learning Process:
Data Preparation that includes attributes selection, cleaning the data, filling the missing values, etc...

Data preprocessing: must use at least 2 methods of data discretization, and compare the final results obtained after each of them.

Learning Proper: for each experiment describe below use a classification tool for rules generation applied to the TWO sets of preprocessed data and compare the results.

Discriminant Rules Generation Experiments: you have to perform 3 experiments (all on the same preprocessed data)

Experiment 1: use all records to find rules for the full classification; i.e. rules describing all classes C1- C6 simultaneously.

Experiment 2: use all records to find rules contrasting class C1 with all others

Experiment 3: repeat Experiments 1, 2 for all records with the most important attributes as defined by the expert only.

Write a detailed Project Description with methods, motivations, results and submit HARD COPY to the Professor