cse634
DATA MINING

Professor Anita Wasilewska

Spring 2018

COURSE SYLLABUS

Course Web Page
www.cs.stonybrook.edu/~ cse634

The webpage contains:

Detailed Lectures Notes slides

Some Course Book slides

Some previous Research Presentations

Course Syllabus

Please check it often- this is also a way I will communicate with you

Course Text Book

DATA MINING

Concepts and Techniques

Jiawei Han, Micheline Kamber

Morgan Kaufman Publishers, 2003,2011

Second Edition

There is a new Third Edition, but we will follow the Second one as it is more widely available (and cheaper)

We will follow the book very closely

Course Description

Data Mining, called also Knowledge Discovery in Databases (KDD) and now called also BIG DATA is a multidisciplinary field

It brings together research and ideas from

database technology,

machine learning,

statistics, pattern recognition, knowledge based systems, information retrieval, high-performance computing, and

data visualization to name the few

Course Description

Data Mining 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

Course **Lectures** are designed to explain in details the material from book chapters

Course Description

The course is designed to give a broad, yet in-depth **overview** of the Data Mining field

It will examine **the most BASIC recognized algorithms and techniques** in a rigorous detail

It also will explore the newest trends and developments of the field in a form of student's talks based on newest research developments and papers from the field - these will be subjects of student's Research Presentations

COURSE STRUCTURE

Part 1 Introduction

Book chapters 1, 2 and Lectures 1, 2

Part 2 Classification

Decision Tree Induction and Neural Networks

Book chapter 6 and Lectures 3 - 7

Team Classification Project

See the Project Description in Syllabus and check the link on the course Website.

COURSE STRUCTURE

Part 3 Association Analysis

Apriori Algorithm

Classification by Association

Book chapter 5 and Lectures 8, 9

Test Review One

Lecture 10

Part 4 Genetic Algorithms

Genetic Algorithms Introduction

Genetic Algorithms Examples

Book chapter 6, Lectures 11, 12

COURSE STRUCTURE

Test Review Two

Lecture 13

Midterm/Final Test

It is in class test and covers material from **Parts 1- 4**

Part 5 Cluster Analysis

Book chapter 7 and Lectures 14

Part 6 Foundations of Data Mining

Lecture 15

Part 7 Students **Research Presentations**

Attention: Project and Research Presentations are to be conducted in **teams**

GRADING COMPONENTS

During the semester students are responsible for the following (in order as listed)

Team Project (40pts)

Midterm/Final Test (70pts)

Team Research Presentation (60pts)

Final Report (30 points)

FINAL GRADE COMPUTATION

NONE of 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 is **translated** into a letter grade in a standard way 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

F is below 60%

Course Contents and Schedule

The course will follow the book very closely and in particular we will cover **all** or **parts**of the following chapters and subjects

The order does not need to be sequential

Chapter 1

Introduction and General overview

What is Data Mining, which data, what kinds of patterns can be mined - Lecture

Chapter 2

Data preprocessing

Data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation

Lecture

Course Contents and Schedule

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

Course Contents and Schedule

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

Course Contents and Schedule

Chapter 7

Cluster Analysis

A Categorization of major Clustering methods

Lecture and students Presentations

Chapters 8-11

Applications and TRENDS in DM

Reading and /or students presentations

Foundations of Data Mining

SPRINGER Encyclopedia of Complexity and Systems Science, 2009 Editors: Editor-in-chief: Meyers, Robert A
http://www.springer.com/us/book/9780387758886

PROJECT DESCRIPTION

Project goal is to use Internet based Classification Tools to build two type classifiers: **descriptive** and **non-descriptive**

Discuss the results in both cases

Compare these two approaches on the basis of obtained results

The detailed project description is in the course Syllabus

It also is published as a link published at the course webpage

PROJECT DESCRIPTION

Descriptive Classifier

Use a Decision Tree tool to generate sets of **discriminant rules** describing the content of the data.

Use WEKA
http://www.cs.waikato.ac.nz/˜ ml/weka/index.html)

Non-Decsriptive Classifier

Use Neural Networks tool to build your classifier

Use WEKA or a tool of your choice

Describe specifics of your tool in a way that makes your report comprehensible for others.

PROJECT DESCRIPTION

Project data is provided on the course web page

This is a real life classification data with TYPE DE ROCHE (Rock Type) as a class attribute

There are 98 records with 48 attributes and 6 classes

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

The project has to follow the steps of **DM Process** to build different classifiers defined by three experiments

Project Experiments

Experiment 1

Use the preprocessed data to perform a **full classification** (learning).

This means build a classifier for all classes **C1- C6** simultaneously

Experiment 2

Use the preprocessed data to perform a **contrast classification** (contrast learning).

This means build a classifier **contrasting** class **C1** with a class **notC1** that contains other classes

Project Experiments

Experiment 3

Repeat Experiments 1, 2 for reprocessed data with the **most important attributes** as defined by the expert

Write a detailed **Project Description** with methods, motivations, results and submit via e-mail to TA and Professor

It is a **team project**

The teams are the same is the for the Research Presentation

Research Presentations

Each **presentation** must consists of the following **two parts**

Part 1 (40pts)

It is a Lecture type, 20 - 25 minutes long presentation

Part 2 (20pts)

It is a short, 5 - 10 minutes presentation of a research paper , or an application

Research Presentation

Presentation Part 1 main goal is to **teach others** the material

It must be a detailed, Lecture type presentation

It can be be based on, or extending the content of the book not covered by course lectures,

It can also cover **other subjects** not covered in the course book and taken from other sources

Research Presentation

Presentation Part 2

It is a presentation of a research paper or of a newest commercial application **connected with** the subject covered in the **Presentation Part 1**

The structure of the **Presentation Parts 1, 2** is described in the Syllabus

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.

Presentation s Subjects

Students can find their own subjects

But here are **suggestions** of some possible subjects

Data Warehouse and OLAP technology - Chapter 3 of the Book

Data Cube Computation and Data Generalization - Chapter 4 of the Book

Presentation s Subjects

Statistical Methods 1

Statistical Prediction, Prediction by Regression, or any other purely statistical methods

Statistical Methods 2 - Bayesian Classification

Statistical Methods 3 - Cluster Analysis and categorization of major Clustering methods

Evolutionary Computing

Genetic algorithms as optimization

Genetic algorithms as classification

Other evolutionary computing methods.

Presentation s Subjects

NEW ADVANCES] in Data Mining

Deep Learning

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

FINAL REPORT

Each student has to write a **report** about 10 research presentations

The detailed format of the report is in the course Syllabus

It os also published as a link published at the course webpage