cse521
DATA MINING

Professor Anita Wasilewska

Spring 2022

COURSE SYLLABUS

Course Web Page
www.cs.stonybrook.edu/~ cse521

The webpage contains:

Detailed Lectures Notes slides

Some course Book slides

Slides of some previous Data Mining Presentations

Course Syllabus

Course Project Description and data

Course Final Report Description

Please **check it** often

Course Text Book

DATA MINING

Concepts and Techniques

Jiawei Han, Micheline Kamber

Morgan Kaufman Publishers, 2003, 2011

Second or Third Edition

My Lectures follow mainly the Second Edition

It is more widely available (and cheaper)

There is no essential differences between the editions

We will follow the book very closely

Course Description

Data Mining 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 BASIC material from book chapters

Course Description

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

We will examine **slowly** and in rigorous detail the most **basic** and **important** algorithms and techniques

We will also will explore the newest trends and developments

GRADING GENERAL PRINCIPLES

1. ALL TESTS will be given as a **Take Home** test

2. ALL TESTS, PROJECT, and FINAL REPORT are to be conducted in **Teams** of 2 - 3 students

3. ALL members of the **Team** receive the same grade

4. NONE of the grades will be curved

TAKE HOME TEST POLICY

TAKE HOME TEST means that you take it at **home** and have **access to**, and can freely **use** the course TEXTBOOK, course Lectures Slides, and all other Presentations posted on the course **Webpage**

You will have one full day to **complete** each TEST
This schedule is designed to give you **time to think** and to write solutions carefully and clearly

TAKE HOME TEST POLICY

Clarity and style of your **solutions** will be important part of your **grade**

Straightforward **copy** of what was published and you have found in the the materials you have **access to** will result in **0pts** for the problem

You always have to write your **solutions** in your own words and to do it in such way as to make it VISIBLE to us that you understood the material

TEAMS FORMATION

ALL TESTS, PROJECT, and FINAL REPORT are to be conducted in **Teams** of 2 - 3 students

Please email **TA** tba names, IDs, and emails of your **Team** members pointing out the Team Leader

TA will assign a Team number to each Team to be used for future **identification**

Contact him if you **do not have** a team partner and he will help you to form a **Team**

This has to be done by **February 24**

COURSE STRUCTURE

The course **Lecture Slides** are written by me, except when I say "Book Slide" or give other credentials

We list here chapters numbers from 2nd edition

We follow 2nd edition chapter numbers by chapters numbers from 3rd edition put between parenthesis

We will follow the **course structure** below

Part 1 Introduction

Data Preprocessing, Data Warehouse

Book chapters 1- 3 (1 - 4) and Lectures 1- 3

COURSE STRUCTURE

Part 2 Classification

Decision Tree Induction and Neural Networks

Book chapter 6 (8- 9) and Lectures 4 - 11

Test 1 Review Lecture

TEST 1

Classification Project

Project Description is posted on the course webpage

COURSE STRUCTURE

Part 3
Association Analysis

Apriori Algorithm

Classification by Association

Book chapters 5, 6 (6, 9) and Lectures 12 - 14

Part 4 Other Classification Methods

Genetic Algorithms

Bayesian Classification

Book chapter 6 (9) Lectures 15, 16

Test 2 Review Lecture

TEST 2

COURSE STRUCTURE

Part 5 Clustering, Statistical Prediction

Book chapter 7 (10, 11) and Lectures 17, 18

Part 6 Other DM Areas and **Foundations of DM**

Chapters 9 - 10 (13) and Lectures 19 -23

We will also cover, if time allows, in some level of detail the following subjects

Types of Neural Networks

Protein Secondary Structure Prediction

Descriptive Granularity - a Data Mining Model

GRADING COMPONENTS

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

1. Test1 (70pts)
2. Classification Project (30pts)
3. Test 2 (70pts)
4. Final Report (30points)

TESTS SCHEDULE

Preliminary Test Schedule

CHANGES WILL BE POSTED ON BLACKBOARD and course webpage

TEST 1 THURSDAY, MARCH 10

Project due March 24

TEST 2 TUESDAY APRIL 23

Final Report due May 7 - last day of classes

TESTS SCHEDULE

TEST 1 is posted MARCH 10 at 12 am and is due MARCH 10 at any time before or at 11:59pm

TEST 2 is posted APRIL 23 at 12am and is due APRIL 23 at any time before or at 11:59pm

There is **no** in class LECTURE on the days of the TESTS

FINAL GRADE COMPUTATION

NONE of GRADES will be CURVED

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 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

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

Course Contents

Chapter 2

Data preprocessing

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

Chapters 3

Data Warehouse

Chapter 5

Mining Association Rules in Large Databases

Transactional databases and Apriori Algorithm

Course Contents

Chapter 6

Classification and Prediction

1. Decision Tree Induction ID3, C4.5

2. Neural Networks

3. Bayesian Classification

4. Classification by Association rule mining

5. Genetic algorithms

6. Statistical Prediction

Course Contents

Chapter 7

Cluster Analysis

A Categorization of major Clustering methods

Chapter 8

Mining Sequential Patters in Biological Data

Chapter 10]

Text Mining

Chapter 11

Foundations of Data Mining