CSE521
DATA MINING

Spring 2023

Course Webpage:

http://www3.cs.stonybrook.edu/˜ cse521

Professor Anita Wasilewska

Meets Tuesday, Thursday 11:30 am - 12:50 pm

Place : Melville Library, Room E4320

Professor Anita Wasilewska

e-mail anita@cs.stonybrook.edu

Office phone (631) 632 8458

Office location: New Computer Science Department, office 208

Professor Office Hours

Tuesday, Thursday 5:00 pm - 6:00 pm and by appointment

Place New Comp. Science Building, Room 208, telephone: 2-8458

In person in the Office and Zoom on demand and in a case of snow emergency

I also read emails DAILY and respond within a day or two to students e-mails

TAs Office Hours will be posted and updated on BLACKBOARD and mailed to all students

TA Office Location Room 2126 Old CS Building

In person in the TAs Office and Zoom on demand and in a case of snow emergency

TAs are **responsible** for tests and assignments grading, Professor writes the tests, assignments and solutions and sets grading criteria and discusses them with TAs

Textbook

DATA MINING Concepts and Techniques

Jiawei Han, Micheline Kamber, and ian Pei Morgan Kaufman Publishers, 2011

Second or Third Edition

Book Slides for Third Edition $http : //hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm$

Course Description

Data Mining (DM), called also Knowledge Discovery in Databases (KDD) is a 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. We also will explore the newest trends and developments of the field. In particular we will cover all or part of the following subjects

Grading General Principles

1. **TEST 1** and **TEST 2** are IN CLASS tests
2. PROJECT, FINAL REPORT PRESENTATION and FINAL REPORT PAPER are to be conducted in **Teams** of 4 - 5 students
3. All members of the **Team** receive the **same grade**
4. **NONE of the grades will be curved**

Teams Formation

Please email TA (to be specified later) names, IDs, and emails of your **Team** members denoting who is the designated **Team Leader**. TA will assign a **Team Number** to each Team and email it to each Team Leader to be used for future identification.

Contact the TA if you **do not have** a team partner. He will help you to FORM a **Team**

Make-up Exams

Make-up exams will be given only in extenuating circumstances (e.g., doctor's note stating that you were ill and unfit to take the exam). Students who miss an exam for a valid reason must contact the instructor immediately to take a make-up exam at the earliest possible time; specific arrangements will be made on a case-by-case basis.

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 followed by Chapters numbers from 3rd edition put between parenthesis

Part 1 Introduction; Data Preprocessing, Data Warehouse

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

Part 2 Classification

Decision Tree Induction and Neural Networks

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

TEST 1

Classification Project Project Description is published at the course webpage

Part 3 Association Analysis, Apriori Algorithm

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

Part 4 Other Classification Models

Genetic Algorithms

Bayesian Classification

Book chapter 6 (9) Lectures 15, 16

TEST 2

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

Final Report Final Report Description is published at the course webpage. It consists of two parts:

Final Report PRESENTATION and Funal Report PAPER

Grading Components

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

1. Test1 (70pts)

3. Project (30pts)

3. Test 2 (70pts)

4. Final Report Presentation (20pts)

5. Final Reports Paper (10pts)

FINAL GRADE COPMUTATION

Attention: **NONE of the 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 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%.

Preliminary Test Schedule

CHANGES WILL BE POSTED ON BLACKBOARD and course webpage

TEST 1   TUESDAY MARCH 7

Spring Break March 13 - 19

Project   due Tuesday, March 27 - submit to Blackboard

TEST 2   THURSDAY APRIL 13

Final Report Presentation   APRIL 18 - MAY 2

Final Report Paper] - due the last day of classes MAY 5 - submit to Blackboard

Course Contents

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

Chapters numbers below are from 2nd edition. Respective Chapters numbers in 3rd edition are listed in the **Course Structure** section.

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.

Chapter 3 Data Warehouse and OLAP technology for Data Mining.

Chapter 5   Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm

Chapter 6 Classification and prediction.

1. Decision Tree Induction ID3, C4.5).

2. Neural Networks

3. Bayesian Classification

4. Classification based on Concepts from Association rule mining

5. Genetic algorithms

6. Statistical Prediction

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 and also in

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

Make-up Exams Make-up exams will be given only in extenuating circumstances (e.g., doctor's note stating that you were ill and unfit to take the exam). Students who miss an exam for a valid reason must contact the instructor immediately to take a make-up exam at the earliest possible time; specific arrangements will be made on a case-by-case basis.

Stony Brook University Syllabus Statement If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact Disability Support Services at (631) 632-6748 or http://http://studentaffairs.stonybrook.edu/dss They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential.

Students who require assistance during emergency evacuation are encouraged to discuss their needs with their professors and Disability Support Services. For procedures and information go to the following website:

http://www.sunysb.edu/ehs/fire/disabilities.shtml

Student Accessibility Support Center Statement

If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, 128 ECC Building, (631) 632-6748, or via e-mail at: sasc@stonybrook.edu. They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential.

Academic Integrity Statement 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. Any suspected instance of academic dishonesty will be reported to the Academic Judiciary. For more comprehensive information on academic integrity, including categories of academic dishonesty, please refer to the academic judiciary website at http://www.stonybrook.edu/uaa/academicjudiciary/

SASC Student Accessibility Support Center

If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, Stony Brook Union Suite 107, (631) 632-6748, or at sasc@stonybrook.edu

Critical Incident Management

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of University Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Faculty in the HSC Schools and the School of Medicine are required to follow their school-specific procedures. Further information about most academic matters can be found in the Undergraduate Bulletin, the Undergraduate Class Schedule, and the Faculty-Employee Handbook.