CSE521  
DATA MINING  
Spring 2022  
Course Webpage:  
http://www3.cs.stonybrook.edu/~cse521  
Professor Anita Wasilewska

Meets  Tuesday, Thursday  1:15pm - 2:35pm

Place  Humanities 1006

Professor  Anita Wasilewska

e-mail  anita@cs.stonybrook.edu

Office phone  (631) 632 8458

Office location:  New Computer Science Department, office 208

Office Hours:  Tuesday, Thursday  11:30 am - 12:30 pm item[] In person in the Office and /or ZOOM

ZOOM  tba

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

Teaching Assistants  tba on course webpage

TAs office hours:  tba

TA e-mail  and/or zoom to be announced

TA Office Location  tba

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  
Morgan Kaufman Publishers, 2003  
Second or Third Edition
Web page:  https://han.j.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. 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 Tests Policy
TAKE HOME TEST means that you take it at home and have access to, and can freely use the BOOK, Lectures Slides and Presentations and all information posted on your course Web Page.
You also will have one full day to complete the tests. This schedule is designed to give you time to think deeper and to work on your tests problems longer. It gives you time to write solutions carefully and clearly. 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 - as in any case of cheating.
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
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
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

Grading Components

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

1. Test1 (70pts)
2. Project (30pts)
3. Test 2 (70pts)

FINAL GRADE COMPUTATION

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:  \( \frac{\# \text{ of earned points}}{2} = \% \text{ grade} \).

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  THURSDAY, MARCH 10
Spring Break  March 14 - 20

Project  due Thursday, March 24 - submit to Blackboard

TEST 2  TUESDAY APRIL 23

Final Report  due last day of classes: May 7 - submit to Blackboard

All TESTS are posted on Blackboard n Assignments by 12 am of the day the TEST and are due any time before, or at 11:59 PM of same the day of the TEST

TEST 1 is posted MARCH 10 at 12am 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

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 Patterns in Biological Data
**Chapter 10** Text Mining
**Chapter 11** Foundations of Data Mining and also in


**Required Syllabi Statements:** The University Senate has authorized that the following required statements appear in all teaching syllabi on the Stony Brook Campus.

**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 Student Conduct and Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students’ ability to learn. Until/unless the latest COVID guidance is explicitly amended by SBU, during Spring 2022 "disruptive behavior will include refusal to wear a mask during classes. (If the syllabus is in print only, then add:) For the latest COVID guidance, please refer to: https://www.stonybrook.edu/commcms/strongertogther/latest.php

**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.