CSE634 Data Mining, Spring 2019
Professor Anita Wasilewska

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

Meets Tuesday, Thursday 4:00 pm - 5:20 pm
Place Engineering 145

Professor Anita Wasilewska
  e-mail address: anita@cs.stonybrook.edu
  Office phone number: 632 8458
  Office location: New Computer Science Department, Room 208

Office Hours Tuesday, Thursday 5:30 pm - 7:00 pm, and by appointment.

TA t.b.a

There will be multiple TAs

TAs are responsible for grading

All grades are listed on BLACKBOARD

TAs office hours and other responsibilities will be listed on the course webpage

Textbook
  DATA MINING Concepts and Techniques
  Jiawei Han, Micheline Kamber
  Morgan Kaufman Publishers, 2003
  Second or Third Edition

Book Slides for Third Edition


Course Description
  Data Mining (DM), 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. We
  also will explore the newest trends and developments of the field. In particular we will cover all or
  part of the following subjects
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 4-8

Midterm 1 Review
Midterm 1

Classification Project
To see the Project Description check the link on the course webpage

Part 3  Association Analysis
Apriori Algorithm
Classification by Association
Book chapters 5, 6 (6, 9) and Lectures 9, 10

Part 4  Other Classification Models
Genetic Algorithms
Bayesian Classification
Book chapter 6 (9) Lectures 11, 12, 13

Midterm 2 Review
Midterm 2

Part 5  Clustering, Statistical Prediction
Book chapter 7 (10, 11) and Lectures 14, 15

Part 6  Other DM Areas and Foundations of DM
Chapters 9 - 10 (13) and Lectures 16, 17

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**  The Final Report description will be published at the course webpage

**Attention**  Project and Final Report can to be conducted in **Teams**

Team can consists of 3 - 6 students and must be the SAME for Classification Project and Report. All members of the Team receive the same grade.

You have to submit **Team** members names to a designated TA to be advertised on the course webpage

**Grading Components**

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

1. Midterm 1 (70pts)
2. Midterm 2 (70pts)
3. Project (30pts)

**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:  \# 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**

Midterm 1  Tuesday, MARCH 12 in class

Spring Break  March 18- 24

Project  due March 26 - submit to Blackboard

Midterm 2  Tuesday, APRIL 23 in class
Final Report  due May 9 - 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 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.

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.