CSE 5710-01: Data Mining for Data Science

Spring 2023

February 19, 2023

1 Overview

Description: This course presents an introduction to data mining algorithms in the areas of classification, association analysis, clustering, and anomaly detection, with an emphasis on a conceptual understanding these algorithms along along with their application in real-world problems and domains.

Pre-requisites: Open to students in the Unviersity Master of Science in Data Science Program. CSE 5709 - Machine Learning for Data Science

Possible Updates to this Document: Excluding materials for purchase, syllabus information may be subject to change.

Meeting Times and Location: Tuesday, Thursday 2:00 - 3:15 pm, ROWE 213

Instructor: Joe Johnson (joseph.2.johnson@uconn.edu)

Instructor Office Hours: Tuesdays, 4-5:30 pm.

Textbook Required: *Introduction to Data Mining, 2nd Edition.* Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Pearson. 2019. ISBN-13: 978-0133128901. ISBN-10: 0133128903.

Recommended Reference 1: *Introduction to Statistical Learning with Applications in R.* Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2013. ISBN-13: 978-1461471370. ISBN-10: 1461471370. Freely accessible on the web at http://faculty.marshall.usc.edu/gareth-james/ISL/index.html.

Recommended Reference 2: *Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition.* Ian H. Witten, Eibe Frank, Mark A. Hall and Christopher J. Pal. Morgan Kaufmann. 2017. ISBN-13: 978-0128042915. ISBN-10: 0128042915.

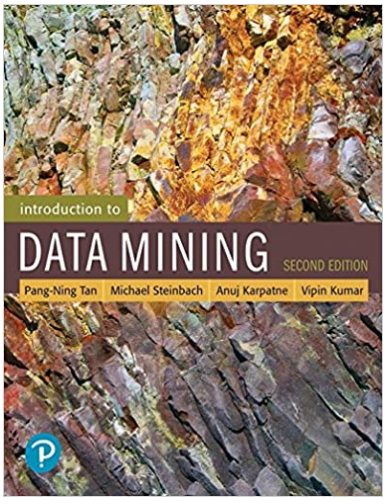

Figure 1: Required Text: Introduction to Data Mining

Course Component	Grade
Homework Assignments	30%
Programming Assignments	30%
Midterm Exam	20%
Final Exam	20%

Table 1: Course Components and Grade Weightings

Communication: The principal means of communication are as follows:

- In-class lectures. See scheduled meeting times/locations, above.
- Announcements in HuskyCT. (Be sure to monitor your email as any announcements will be forwarded to your email address.)
- Email: joseph.2.johnson@uconn.edu
- Virtual Office Hours: Tuesday 4-5:30 pm

Prerequisite Courses: Background in statistics at the undergraduate level, or permission of the instructor. Reasonable proficiency in the Python programming language is required.

Course Approach:

- This is a 15-week course, where each week opens on Monday morning and ends at 11:59 pm the following Sunday.
- Each week, new material will be posted on the HuskyCT Course Content folder, labeled Lecture XX – YYY, where XX corresponds to the week of the course and YYY is the title/topic for that week's lecture.
- Assignment work will be due 11:59 pm on the Sunday of the week the assignment is due.

Grading, Assignments and Exams There will be approximately 4 homework sets, 3 programming assignments, and 2 exams throughout the semester. However, these assessments are subject to change as circumstances dictate. The weights toward the final course grade are given in Table 1. The exact number of homework and programming assignments may vary based on how problems are grouped into assignments as we progress through the material. **These assignments are to be submitted on HuskyCT, as a pdf document.** The problem sets can be typed or handwritten and the programming assignments must be typed (as Python code in Jupyter Notebooks, and then stored, submitted as pdf files).

Late Policy for Submitting Assignments

- Homework is to be handed in by 11:59 Eastern Time on the day it is due.
- There will be a 3 hour grace period after this time so that if you hand it in by 2:59 am of the following day, there will be no penalty.
- After the grace period, the assignment is considered one day late, for which there will be 10 point penalty.
- For each 24 period hour interval after the 2:59 am grace period, the assignment will be considered another day late, with a 10 point deduction for each day.
- After one week late, the assignment will not be graded.

Number Grade	Assigned Grade	Grade Points
90-100	A	4.0
80-89	B	3.0
70-79	C	2.0
60-69	D	1.0
0-59	F	0.0

Table 2: Grading Scale

Collaboration All homework assignments must be completed individually. It is okay for you to discuss a problem with a classmate as long you abide by the following condition:

- Each student you collaborate with should be named on the homework assignment.
- You must first consider each problem on your own and generate ideas on how to solve the problem.
- You may discuss problems and ideas jointly. The goal of collaboration is to understand the high level ideas of the problem. Do not go further than this.
- You must write solutions completely on your own.
- The midterm exams and the final exam will be an individual effort.
- Do not use other resources (outside of your textbooks and collaborators) to attempt to find the problem or the solution. This includes using the internet to search for parts of the problem.

2 Course contents

Course Objectives By the end of this course, you will be able to:

- Discuss and apply various data preparation and integration techniques, including aggregation, sampling, dimensionality reduction, feature extraction, and variable transformation.
- Discuss the theoretical underpinnings of the various data mining approaches, including notions of vectorization, measures of vector distance and similarity, entropy.
- Explain the theoretical principles of each the various data mining approaches discussed in the course, including but not limited to linear regression, logistic regression, support vector machines, decision trees, k-means, db-scan, neural networks. Compare and contrast the various approaches in data mining.
- Use state-of-the-art tools to apply, evaluate, and compare the performance of various data models.
- Apply data preparation and mining techniques to construct powerful predictive models in scientific and industrial settings.

Tentative Class Schedule Table 3 contains a tentative plan for the topics by week along with target dates for homework assignments, programming assignments, and exams. As this course is intended to serve as a kitchen sink into which we toss a whole array of topics, this list is subject to change.

3 Policies

Academic Honesty This course expects all students to act in accordance with the Guidelines for Academic Integrity at the University of Connecticut. If you have questions about academic integrity or intellectual property, you should consult with your instructor. Additionally, consult UConn's guidelines for academic integrity.

The collaboration policy described above is designed to allow students the resources to succeed while ensuring they learn and master the material. If you are unsure if something is acceptable according to the collaboration policy, talk to me!

Violations of this policy will be considered violations of the academic integrity policy and will be reported to the Academic Integrity Hearing Board. Consequences may include (but are not limited to) failure of the class. Example violations include: not reporting collaborators, jointly writing solutions, copying or plagiarizing solutions from other sources, and cheating on the exam.

Week	Date	Topics	Tan Chapters	Exam	Assignment
1	01/17	Lecture 01: Intro to KDD and Data Mining	1		HW1: 01/10-01/22
2	01/23	Lecture 02: Data and Data Preparation: Data Types, Data Quality, Data Preprocessing, Measures of Similarity and Dissimilarity	2		
3	01/30	Lecture 03: Classification - Part 1: Basic Concepts, General Framework, Decision Tree Classifier	3.1-3.3		PA1: 01/23-02/05
4	02/06	Lecture 04: Classification - Part 2: Model Overfitting, Model Selection, Model Evaluation, Hyperparameters, Pitfalls of Model Selection and Evaluation, Model Comparison	3.4-3.9		
5	02/13	Lecture 05: Classification Techniques - Part 1: Types of Classifiers, Rule-Based Classifiers, Nearest Neighbor Classifiers, Naive Bayes Classifier, Bayesian Networks	4.1-4.6		HW2: 02/06-02/19
6	02/20	Lecture 06: Classification Techniques - Part 2: Support Vector Machines (SVMs), Ensemble Methods, Class Imbalance Problem	4.9-4.11		
7	02/27	Lecture 07: Numeric Prediction: Linear Regression, Regression Trees, Model Trees	ISLR: Ch. 3		PA2: 02/20-03/05
8	03/06	**Midterm Exam Week**		**Midterm:** 03/06-03/08	
	03/13	**Spring Break**			
9	03/20	Lecture 08: Association Analysis: Frequent Itemset Generation, Rule Generation, Compact Representations of Frequent Itemsets, Alternative Methods	5.1-5.5		HW3: 03/13-03/26
10	03/27	Lecture 09: Classification Techniques - Part 3: Logistic Regression, Artificial Neural Networks (ANNs), Deep Learning	4.7-4.8		
11	04/03	Lecture 10: Clustering Analysis: Overview, K-Means, Agglomerative Heirarchical Clustering, DBSCAN, Cluster Evaluation	7		PA3: 03/27-04/09
12	04/10	Lecture 11: Anomaly Detection: Problems and Methods, Statistical Approaches, Proximity-Based Approaches, Clustering-Based Approaches, Reconstruction-Based Approaches, Evaluation of Anomaly Detection Methods	9		
13	04/17	Lecture 12: Avoiding False Discoveries: Preliminaries, Modeling Null and Alternative Distributions, Statsitcal Testing for Classification, Association Analysis, Cluster Analysis, Anomaly Detection	10		HW4: 04/10-04/23
14	04/24	Review Week			
15	05/01	**Final Exam Week**		**Final Exam:** 04/24-04/26	

Table 3: Tentative class schedule

Student Conduct Code Students are expected to conduct themselves in accordance with UConn's Student Conduct Code.

Students with Disabilities The University of Connecticut is committed to protecting the rights of individuals with disabilities and assuring that the learning environment is accessible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, please let me know immediately so that we can discuss options. Students who require accommodations should contact the Center for Students with Disabilities, Wilbur

Cross Building Room 204, (860) 486-2020, or http://csd.uconn.edu/.