CSCI 5523: Introduction to Data Mining

Syllabus

Units: 3

Term — Day — Time: Spring 2023, MW – 2:30-3:45 pm

Location: Keller Hall 3-210
Website: TBD

Instructor: Yao-Yi Chiang, Ph.D. GISP
Office: Keller Hall 519
Regular Office Hours: Wednesdays after class
Contact Info: yaoyi@umn.edu

Teaching Assistant: Yijun Lin, TBD
Office Hours: TBD
Contact Info: lin00786@umn.edu

1. Course Description

Data mining is foundational to the data analytics skill set. At a high level, it allows the analyst to discover patterns in data and transform them into a usable product. The course will teach data mining algorithms for analyzing very large data sets. It will have an applied focus in that it is meant to prepare students to utilize topics in data mining to solve real-world problems.

2. Recommended Preparation

Knowledge of probability, linear algebra, data structure, algorithms, and machine learning.

A basic understanding of engineering principles is required, including good programming skills; familiarity with Python is desirable. Most assignments are designed for the Unix environment; basic Unix skills will make programming assignments much easier. Students need sufficient mathematical background, including probability, statistics, and linear algebra. Some knowledge of machine learning is helpful but not required.

3. Textbook

Rajaraman, J. Leskovec and J. D. Ullman, Mining of Massive Datasets. Cambridge University Press (3rd Edition). Available free at: http://www.mmds.org/

5. Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings, and students are encouraged to do the readings before the discussion in class. All course materials will be available online, including the readings, lecture slides, and homework.

6. Technological Proficiency and Hardware/Software Required

Students are expected to know how to use common operating systems (e.g., MS Windows and Mac OC) and install applications on the operating systems.

Students are also expected to have their laptop or desktop computer where they can install and run software to finish the homework assignments.

7. Description and Assessment of Assignments

Homework Assignments: There will be 6 programming assignments based on the class topics each week. The assignments must be done individually. Each assignment is graded on a scale of 0-100, and the specific rubric for each assignment is given in the assignment. The assignments must be done individually. Each assignment is graded on a scale of 0-100, and the specific rubric for each assignment is given in the assignment. Each submission will be checked for plagiarism.

Quizzes: There will be weekly quizzes based on the material from the week before.

Exams: There will be one comprehensive exam on the last day of the class covering all of the material covered in the class (no midterm, no final exams).

Data Mining Competition Project: There will be a final project based on the topics introduced in class. The final project is to build an advanced recommendation system and compete with other students and TAs to achieve the lowest recommendation errors. The project serves as the final summarization purpose of this course and is due on the scheduled University final exam day.

Grading Breakdown

There are two necessary conditions for passing this class: (1) submission of all assignments and (2) scoring at least 50% on the comprehensive examination.

Grading Schema:

Quizzes	30%
Homework	48%
Comprehensive Exam	17%
Data Mining Competition Project	5%
Total	100%

Grades will range from A through F. The following is the breakdown for grading:

94 – 100 = A 74 – 77 = C

90 – 94 = A- 70 –74 = C-

87 – 90 = B+ 67 – 70 = D+

84 – 87 = B 64 – 67 = D

80 – 84 = B- 60 – 64 = D-

77 – 80 =C+ Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 11:59 pm CT on the due date and should be submitted on Canvas. Late submissions within 24 hours of the due date will receive a 30% penalty. Late submissions after 24 hours of the due date will receive a 70% penalty.

Every student has FIVE free late days for homework assignments. You can use these free late days for any reason, separately or together, to avoid the late penalty. There will be no other extensions for any reason. **You cannot use the free late days after the last day of the class.**

8. Schedule

Week	Topics	Readings	Assignments/Exams
1 1/16* Monday 1/16 is	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File	

university holiday		Systems and Map-Reduce	
2 1/23	MapReduce (cont.)	Ch2: Large-Scale File Systems and Map-Reduce	
3 1/30	Frequent itemsets and Association rules	Ch6: Frequent itemsets	HW1 published
4 2/6	Frequent itemsets and Association rules	Ch6: Frequent itemsets	
5 2/13	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	HW1 due, HW2 published
6 2/20	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	
7 2/27	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems, additional readings	HW2 due, HW3 published
9 3/13	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems, additional readings	
10 3/20	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	HW4 published
11 3/27	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	HW3 due, Competition Project Assigned
12 4/3	Clustering	Ch7: Clustering	HW4 due, HW5 published
13 4/10	Mining data streams	Ch4: Mining data streams	
14 4/17	Link Analysis	Ch5: Link Analysis	HW6 published
15 4/24	Web Advertising	Ch8: Advertising on the Web	HW5 due

16 5/1	Comprehesive Exam		Homework 6 due,
		Competition project due 5/7	

9. Modality Transparency:

This course is scheduled as an in-person course. I intend to hold all class sessions in person except if situational factors arise, such as the personal illness of the instructor, when the class may be held synchronously via Zoom or recorded for later viewing.

10. Statement on Academic Conduct and Support Systems

Academic Conduct[1]

You are expected to do your own academic work and cite sources as necessary. Failing to do so is scholastic dishonesty. Scholastic dishonesty means plagiarizing; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using course materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; altering, forging, misrepresenting or misusing a University academic record; or fabricating or falsifying data, research procedures, or data analysis. (Student Conduct Code.) If it is determined that a student has cheated, the student may be given an "F" or an "N" for the course, and may face additional sanctions from the University.

The Office for Community Standards has compiled a useful list of Frequently Asked Questions pertaining to scholastic dishonesty.

Beware of websites that advertise themselves as being “tutoring websites.” It is not permissible to upload any instructor materials to these sites without their permission or copy material for your own homework assignments from these various sites.

If you have additional questions, please clarify with your instructor for the course. Your instructor can respond to your specific questions regarding what would constitute scholastic dishonesty in the context of a particular class, e.g., whether collaboration on assignments is permitted, requirements and methods for citing sources if electronic aids are permitted or prohibited during an exam.

Support Systems

A number of student services can be found on: https://onestop.umn.edu/ and https://disability.umn.edu/.

[1] This is a direct copy from https://communitystandards.umn.edu/syllabus-insertion

11. Legitimate (i.e., Excused) Absences

While **makeup work for legitimate absences** is part of University policy, faculty and instructors choose how to accommodate absences based on their course. In this course, excused absences will be handled as follows:

- Students are expected to obtain notes from a classmate of class material missed.
- Students may request a make-up assignment for class material missed.
- Please note that I do not intend to record class sessions at the request of individual students.
- We will drop your four lowest quiz scores, so we will not have makeup quizzes.

COVID-19 Symptoms, Vaccination, Excused Absences, and Face Coverings

You should stay at home if you experience any signs of illness or have a positive COVID-19 test result. If this occurs, please consult with your healthcare provider about an appropriate course of action. I will follow these same protocols and will let you know if the delivery of this course has to be temporarily changed as a result of my own circumstances. Absences related to illness, including COVID-19 symptoms, for yourself or your dependents, are **legitimate "excused" absences**

Vaccines: COVID-19 Vaccinations (or approved exemptions) are required for all students and employees. Learn about vaccine and booster appointments on campus by visiting the FAQ on Get the Vax page.

Face coverings: Up-to-date policy information is available on the Safe Campus page. The University expects all community members to respect those who choose to wear a mask, as well as those who choose not to wear one.

I don't intend to wear a mask in class myself, and I fully support your individual choices around masking.

Indoor masking continues to be an important tool in high-risk situations. High-quality masks (N-95 or certified KN-95) will be available to students in Fall 2022. Check the Safe Campus website for information on the location(s) for each campus.

Testing: Information on *When*, *Where*, and *What if* for testing is available on MTest webpage.

The above policies and guidelines are subject to change. The University regularly updates **pandemic guidelines** in response to guidance from health professionals and in relation to the prevalence of the virus and its variants in our community.