SI 671/721
Data Mining: Methods and Applications
Fall 2022

Instructor: Paramveer Dhillon (dhillonp@umich.edu).

Time (T) and Location (L):

- **Lecture:**
  (T) Tuesdays 8:30am-10:00am ET,
  (L) 1200 CHEM (Chemistry Building).
- **Discussion Section (DIS 004):**
  (T) Wednesdays 2:30pm-4:00pm ET,
  (L) 2600 SKB (School of Kinesiology Building (formerly the Kraus Building)).
- **Discussion Section (DIS 005):**
  (T) Wednesdays 5:30pm-7:00pm ET,
  (L) 1460 MH (Mason Hall).
- **Discussion Section (DIS 002):**
  (T) Thursdays 8:30am-10:00am ET,
  (L) 2245 NQ (North Quad).
- **Discussion Section (DIS 003):**
  (T) Thursdays 10:00am-11:30am ET,
  (L) 2245 NQ (North Quad).
- **Discussion Section (DIS 006):**
  (T) Thursdays 2:30pm-4:00pm ET,
  (L) 2245 NQ (North Quad).

We will have weekly lectures. The course also has five weekly discussion sections (run by GSIs). Each student is required to attend one discussion section in addition to the weekly lecture. You should have already been assigned to one of the discussion sections when you enrolled for the course. Unfortunately, changing sections is hard since we are limited by the capacity of the sections.

Office Hours:

- Tuesdays, 1:00pm-2:00pm ET *(In-person at my office 3389 North Quad).*
  Zoom link: https://umich.zoom.us/my/dhillonp

GSIs:

- Zihan Wu(ziwu@umich.edu)
  Office Hours:
  - Wednesday 4:15 - 5:15 PM (in person NQ 1270)
  - Friday 4:00 - 5:00 PM (remote) (Zoom Link: https://umich.zoom.us/j/94590155215)
- Piyush Singh (piyushps@umich.edu)
  Office Hours (Remote):
  - Thursdays 12 - 1 PM (in person NQ 1270)
  - Thursdays 1 - 2 PM (remote) (Zoom Link: https://umich.zoom.us/j/2734270593)
- Naren Doraiswamy (narend@umich.edu)
  Office Hours (Remote):
  - Fridays 10AM-12PM (Zoom Link: https://umich.zoom.us/j/93693007703)

Course Website: We'll be using Canvas for disseminating course materials and Slack for communication. You should have been automatically added to both Canvas and Slack workspaces. Here is the link for the Slack workspace in case you haven't joined it as yet um-fa22-si671-721.slack.com

1 Course Description

SI 671/721 is a graduate-level course on advanced topics in data mining. The course provides an overview of recent research topics in the field of data mining, state-of-the-art methods to analyze different types of datasets, and their applications to many real-world problems. The course will highlight the practical applications of data mining instead of the theoretical foundations of machine learning and statistical computing. So, this course can be thought of as a complement to other statistical inference or machine learning courses, but not as a substitute for them.

The course materials will focus on how the information in different real-world problems can be represented as particular genres, or formats of data, and how the basic mining tasks of each genre of data can be accomplished using the state-of-the-art techniques. To this end, the course is not only suitable for masters/doctoral students who are doing research in data mining related fields, but also for students who are consumers of data mining techniques in their own disciplines, such as natural language processing, network science, human computer interaction, biomedical and health informatics, economics, social computing, sociology, and business intelligence.

2 Prerequisites

- Familiarity with Python (and libraries such as Numpy, Pandas, Scikit Learn) will be helpful. If you have a lot of programming experience but in a different language (e.g. R/C++/Matlab), you will probably be fine.

- Basic Probability and Statistics.
- It is recommended that students have some exposure to some basic ideas in statistical learning and optimization via prior courses in related areas such as Machine Learning, Data Mining, Statistical Inference, Econometrics, NLP, Computer Vision, Information Retrieval, Data Science, Social Network Analysis etc.

3 Learning Objectives

Learning objectives of this course include:

- Describe the basic principles of knowledge discovery from data.
- Perform the basic computational tasks of data mining, including pattern and association extraction, data modeling, classification, clustering, ranking, prediction, outlier detection, etc.
- Demonstrate how information in real world applications can be formulated and represented as different genres of data, such as matrices, item sets, sequences, time series, data streams,graphs/networks, etc.
- Identify how to select appropriate data mining techniques for real-world scenarios.
- Identify the major data mining problems specific to different genres of data.
- Apply the state-of-the-art data mining techniques that solve these problems.
- Discuss the various applications of these techniques in multiple disciplines.
- Develop software development skills to deal with large-scale datasets (e.g., at least millions of data records).

4 Readings/Textbooks

There are no required textbooks for this course. Our readings will be derived from research papers published in top conferences/journals in data mining and allied areas such as KDD, WWW, ICDM, CIKM, WSDM, ICWSM, TKDE, TKDD, SIGIR. Here's a list of optional textbooks that can be used for supplemental reading.

1. Leskovec, Rajaraman, Ullman. **Mining of Massive Datasets** (*Much of the content is available online for free:* http://mmds.org/#book)
2. Hastie, Tibshirani, Friedmann. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** (*Available for free online:* https://web.stanford.edu/~hastie/Papers/ESLII.pdf)

5 Course Format and Grading

The course will be instructor-led discussions of different data mining and data science topics. There will also be some GSI-led sessions which will cover code examples. They will be announced one week prior.

5.1 Grading

Grading will be based on:

- Mid-term exam **(20%)**
- Quiz **(5%)**
- Three Programming Assignments. **(45%, 15% each)**
- Final Project. **(30%)**
- Extra Credit: **2%** for answering other students' questions on Slack and forwarding the discussion on a topic.
- Conversion between percentage grades and letter grades will use the following mapping: A+ 100; A 93; A- 90; B+ 87; B 83; B- 80; C+ 77; C 73; C- 70; D+ 67; D 63; D- 60; E 50; F 40.

5.1.1 Mid-term Exam (20%):

There will be an in-class mid-term exam on 10/11. The exam will be closed book + closed notes and will test your knowledge of the concepts covered till the previous lecture (10/4). The exam questions will mostly be multiple-choice.

5.1.2 Quiz (5%):

There will be one 30-minute short quiz on the final day of the class (12/6). It will contain multiple-choice questions based on the second half of the course (after the mid-term).

5.1.3 Programming Assignments (45%):

Students will be required to complete 3 data mining assignments. Each assignment requires students to program different data mining algorithms to solve real-world problems/tasks e.g. Link prediction in social networks; sentiment classification; community detection etc.

5.1.4 Course Project (30%):

A course project is required. Individual projects are preferred but students are allowed to work in teams of 2 or 3. Grading will be adjusted accordingly, so teams of 2 or 3 will have to do proportionately more work. Course project is intended to be an open-ended project based on the interest of student(s). Typically, it will involve developing and implementing

some data mining algorithms for a relevant problem. The project quality should be worthy of being published at a data mining conference. More details will be provided regarding this in the class.

The final deliverables include the software, a detailed write-up, and a presentation at the UMSI Fall exposition in early December. The grading for the course project will be split as follows (of the 30% total):

- Project Proposal (10%): A **concrete** two-page proposal, describing the project topic, objectives, expected deliverable (software package, demo, and/or a technical report), and a list of team members and their expected contribution to the project.
- Final project presentation at Poster-session (10%): Presenting a poster based on your project at a virtual poster session. There might also be other UMSI faculty present at the poster session.
- Final project deliverable and report (10%): Students are expected to submit their project deliverable, along with a detailed report. The report should include key observations and conclusions based on the project and suggest potential follow-up studies. Teams working on the project together must also describe individual contributions of the team members.

6 Weekly Class Schedule

This tentative schedule gives an overview of the topics covered each week in class. Information regarding each week's readings and homework assignments will be made available on Canvas.

- **Week 1 (8/30)**: (Lecture + Discussion) Introduction to Data Mining (I).
- **Week 2 (9/6):** (Lecture + Discussion) Introduction to Data Mining (II).
- **Week 3 (9/13):** (Lecture + Discussion) Mining Itemsets. (**HW 1 is Out. Due 9/29)**.
- **Week 4 (9/20):** (Lecture + Discussion) Mining Matrix Data.
- **Week 5 (9/27):** (Lecture + Discussion) Mining Sequence and Text Data.
- **Week 6 (10/4):** (Lecture + Discussion) Mining Time-series Data (I). (**HW 2 is Out. Due 10/27)**
- **Week 7 (10/11):** (Exam + No discussion section) In-class Mid-term exam (Closed book + Closed notes)
- **Week 8 (10/18):** Fall Break (No Class)
- **Week 9 (10/25):** (Lecture + Discussion) Mining Network Data. (**HW 3 is Out. Due 11/17)**
- **Week 10 (11/1):** (Lecture + Discussion) Mining Streaming Data. (**Final Project Proposal Due)**
- **Week 11 (11/8):** (Lecture + Discussion) Mining Embedded Representations.

- **Week 12 (11/15):** (Lecture) Mining Time-series Data (II).
- **Week 13 (11/22):** Thanksgiving Break (No Class)
- **Week 14 (11/29):** (Lecture) Mining from Experiments: Bandits and Causal Inference.
- **Week 15 (12/6):** (Lecture) Mining interpretable models & Data Science Ethics. (45 minutes) + Short Quiz (30 minutes)
- **UMSI Fall Exposition (12/9): Final Project Poster Presentation**
- **Hard Deadline 12/13: Final Project Report Due**

7 Policies

7.1 Attendance

Attendance at lectures is necessary to get the most out of the course. However, if you're ill you may skip the class and watch the live recorded video of the class.

7.2 Late submission policy

Students have 72 hours of buffer grace period for the entire semester. If necessary, students may use it to submit any of the 3 homework assignments late. A student may use it all on one assignment or use a bit of it for any number of assignments. Once the buffer grace period is used up, late submissions will not be graded. **Please note that the grace period can ONLY be used on the homeworks and NOT on the final project.**

7.3 Academic Conduct

- **Collaboration:** UMSI strongly encourages collaboration while working on some assignments, such as homework problems and interpreting reading assignments as a general practice. Active learning is effective. Collaboration with other students in the course will be especially valuable in summarizing the reading materials and picking out the key concepts.You must, however, write your homework submission on your own, in your own words, before turning it in. If you worked with someone on the homework before writing it, you must list any and all collaborators on your written submission. Each course and each instructor may place restrictions on collaboration for any or all assignments. Read the instructions careful and request clarification about collaboration when in doubt. Collaboration is almost always forbidden for take-home and in class exams.
- **Plagiarism:** All written submissions must be your own, original work. Original work for narrative questions is not mere paraphrasing of someone else's completed answer: you must not share written answers with each other at all. At most, you should be

working from notes you took while participating in a study session. Largely duplicate copies of the same assignment will receive an equal division of the total point score from the one piece of work. You may incorporate selected excerpts, statements or phrases from publications by other authors, but they must be clearly marked as quotations and must be attributed. If you build on the ideas of prior authors, you must cite their work. You may obtain copy editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another. See the (Doctoral, MSI, BSI) student handbooks available on the UMSI intranet for the definition of plagiarism, resources to help you avoid it, and the consequences for intentional or unintentional plagiarism.

- **Reasonable accommodations:** The University of Michigan recognizes disability as an integral part of diversity and is committed to creating an inclusive and equitable educational environment for students with disabilities. Students who are experiencing a disability-related barrier should contact Services for Students with Disabilities https://ssd.umich.edu/; 734-763-3000 or ssdoffice@umich.edu). For students who are connected with SSD, accommodation requests can be made in Accommodate. If you have any questions or concerns please contact your SSD Coordinator or visit SSD's Current Student webpage. SSD considers aspects of the course design, course learning objects and the individual academic and course barriers experienced by the student. Further conversation with SSD, instructors, and the student may be warranted to ensure an accessible course experience. The instructional team will treat any information that you provide in as confidential a manner as possible.

- **Student Mental Health and Wellbeing:** The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources. For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/.

  University Students may experience stressors that can impact both their academic experience and their personal well-being. These may include academic pressures and challenges associated with relationships, mental health, alcohol or other drugs, identities, finances, etc. If you are experiencing concerns, seeking help is a courageous thing to do for yourself and those who care about you. If the source of your stressors is academic, please contact me so that we can find solutions together. Ashley Evearitt, a Counseling and Psychological Services (CAPS) counselor, is embedded in UMSI, information about how to schedule an appointment with her can be found here.

For personal concerns, U-M offers a variety of resources, many which are listed on the Resources for Student Well-being webpage. You can also search for additional well-being resources on that website.

- **Audio and Video Recording of Lectures:**

  Class Recordings: We will be doing audio and video recording of all sessions to enable those who cannot attend class in person on a given day to access the content. These recordings will not be made available publicly. Recordings of all sessions will be available on Canvas only to students registered for this class. As part of your participation in this course, you may be recorded. If you do not wish to be recorded, please contact the professor during the first week of class to discuss alternative arrangements. The camera only picks up the front of the room (instructor and slides), but this may require you to sit in a particular place in the room, outside the cameras' view. Students may not copy and share the lecture videos with those not in the class, or upload them to any other online environment (this is a violation of the Federal Education Rights and Privacy Act (FERPA)).

  Personal recordings are prohibited except with permission: Students are prohibited from recording/distributing any class activity without written permission from the instructor, except as necessary as part of approved accommodations for students with disabilities. Any approved recordings may only be used for the student's own private use.

- **COVID-19 Resources:** Here's a link to Rackham's Resource guide for COVID-19. https://docs.google.com/document/d/1dXbEvnY7_MwkiNPoe_2bzKgGpZdrkxHu5nmMniOulEo/edit.

- **COVID-19 contigency plans:**

  - As per university policy, masking is optional in classrooms and during office hours.
  - If the instructor or GSI is ill or quarantined you will be notified regarding the additional classes/sections to make up for the lost time.
  - Students who are ill and/or quarantined can watch the recorded lectures which will be made available via Canvas. Such students can also miss the in-person discussion sections and watch the recordings.