SI 671/721
Data Mining: Methods and Applications
FALL 2020

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

Time and Location: Mondays 1:00-4:00pm ET, Remote (via Zoom).
https://umich.zoom.us/j/95469645767
The course will have synchronous/Live lectures that will be recorded for students in different time-zones.

Office Hours:

- Monday, 4:15-5:15pm ET (Remote via Zoom)
  https://umich.zoom.us/my/dhillonp
- Tuesday, 9-10am ET (Remote via Zoom)
  https://umich.zoom.us/my/dhillonp

GSIs:

- Kwame Robinson (kwamepr@umich.edu)
  Office Hours (Remote via Zoom): https://umich.zoom.us/my/kwamepr
  - Thursday, 7:15-8:15am ET
  - Friday, 10:30-11:30am ET
- Lingyun Guo (linguo@umich.edu)
  Office Hours (Remote via Zoom): https://umich.zoom.us/my/yunyy
  - Monday, 2:00-3:00am ET
  - Wednesday, 10:00-11:00pm ET

Course Website: We'll be using Canvas for this course.

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:

- 5 quizzes (to be completed asynchronously during the assigned week). **(25%, 5% each)**
- 3 Programming Assignments. **(45%, 15% each)**
- Final Project. **(30%)**
- Conversion between percentage grades and letter grades will use the following mapping: A+ 98; 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 Quizzes (25%):

There will be five multiple-choice question-based quizzes throughout the semester. These quizzes will test the knowledge of the course material covered up till that class. They will be administered remotely, and students will have 24 hours (Monday 4 PM EDT/EST to Tuesday 4 PM EDT/EST) to complete them due to students being in different time-zones.

5.1.2 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.

A sample data mining task used in one of the previous years can be found here: https://inclass.kaggle.com/c/umich-si721-joining-kiva-teams.

5.1.3 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 tentative presentation in a remote poster session in early December. The grading for the course project will be split as follows (of the 40% 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.

  50% of the grade will be given by the instructors and the remaining 50% will be based on double-blind peer review from other students in the class.

- Progress report (5%): A one-page summary of the progress made on the stated project objectives and if there are any hurdles towards timely completion. If there are any significant changes to the submitted proposal, the students should describe them in detail in the progress report. Consider this as a checkpoint towards achieving the stated goals of the project. **There are no penalties for changes to the proposal document, rather it may be more prudent to recalibrate or clarify the expected outcomes during this stage.**

- Final project presentation at Poster-session (5%): Presenting a poster based on your project at a virtual poster session. There might also be other UMSI faculty present at the poster session. This is tentative and might change. If it is not possible to conduct a virtual poster session, then this part of the grade will be added to the final project deliverable described next.

- 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.

  50% of the grade will be given by the instructors and 50% will be based on double-blind peer review from other students in the class.

6 Weekly Class Schedule (Tentative)

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 (8/31)**: Introduction.
- **Week 1.5 (9/3 3:30-5:30PM)**: GSI Recitation: Basics of Applied Data Science.
- **Week 2 (9/7):** Labor Day (No Class)
- **Week 3 (9/14):** Mining Itemsets.
- **Week 4 (9/21):** Mining Matrix Data. **(HW 1 is Out)**
- **Week 5 (9/28):** Mining Time-series Data. **(Project Proposal Due)**
- **Week 6 (10/5):** Mining Sequence and Text Data. **(HW 1 Due)**, **(HW 2 is Out)**
- **Week 7 (10/12):** Mining Streaming Data.
- **Week 8 (10/19):** Mining Network Data. **(HW 2 Due)**, **(HW 3 is Out)**
- **Week 9 (10/26):** Mining User-behavior Data. **(Project Progress Report Due)**
- **Week 10 (11/2):** Mining Image Data.
- **Week 11 (11/9):** Mining Spatio-temporal Data. **(HW 3 is Due)**
- **Week 12 (11/16):** Mining from Randomized Experiments & Interactive Mining.
- **Week 13 (11/23):** Thanksgiving Break (No Class)
- **Week 14 (11/30):** Mining Massive Datasets & Mining interpretable models.
- **Week 15 (12/7) Final Project Presentations.**
- **Finals Week (12/14): Final Project Report Due**

7 Policies

7.1 Attendance

Attendance at lectures is necessary to get the most out of the course. Some of the GSI lab sessions might be optional. Also, everyone is encouraged to turn on their camera while attending the class via Zoom. You need to talk to the instructor in advance if you don't want to. Due to several students being in different time-zones, the Zoom lectures will be recorded and can be watched later. It is okay for those students to watch those lectures at a later time and not attend the synchronous "live" class session.

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.

7.3 Academic Conduct

Collaboration: We strongly encourage 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.

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, MHI, or 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: If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way the course is usually taught may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu.) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

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 mentalhealth services available on campus, please visit:http://umich.edu/~mhealth/.

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.