SI 671 - Data Mining: Methods and Applications

SYLLABUS

Instructor: David Jurgens (jurgens@umich.edu)

Meeting schedule: Mondays, 2:00 – 5:00pm, 1125 North Quad

Office Hour: Thursdays, 9:00 – 11:00 am, 3341 North Quad

With the explosive growth of information generated from different sources, in various formats, and with a variety of characteristics, information analysis has become challenging for researchers in many disciplines. Automatic, robust, and intelligent data mining techniques have become essential tools to handle heterogeneous, noisy, non-traditional, and large-scale data sets.

SI 671 is a graduate-level seminar 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 genres of information, 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. 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 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.

A. Learning objectives

The learning objectives for 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).

- Explore the recent trends and open directions in the field of data mining.

B. Course Format and Grading

The course will be instructor-led discussions of different data mining topics. Class will run the full allotted time. During some weeks, we will do live in-class development on mini-challenges.

B.1. Grading

Grading will be based on:

- One-page summaries and participation in discussion (5 summaries) (10%)
- Mid-term exam (20%)
- Programming assignments (30%; 10% each)
- Course project (40%)

B.1.1. One-page paper response (10%):

Each week, starting week 2, students have the option of writing a one-page response based on the reading assignments for the previous week. A total of five summaries should be submitted, with the student having the option of what to choose. The reading assignments will cover significant papers on the topics being discussed in class. Each one-page response will follow a specific format, shared through Canvas.

B.1.3. Mid-term exam (20%):

The mid-term exam will be a take-home test and will be administered around Week 9 of the semester (early November 2017).

B.1.4. Programming Assignments (30%):

Students will be required to complete in three data mining assignments. Each consists of a written component and using online competition services such as Kaggle-in-Class (http://inclass.kaggle.com). The task will be closely related to the course material, with real-world data and gold-standard judgments provided. Students can submit and resubmit their results to the competition site and get instant feedback (evaluation metrics) from the system. The task will likely to be selected from one of the follows: Link prediction in social networks; sentiment classification; citation prediction; community detection, etc.

Sample Challenge: A sample task of the data mining challenge used in one of the previous years: https://inclass.kaggle.com/c/umich-si721-joining-kiva-teams (Links to an external site.)

B.1.5. Course Project (40%):

A course project is required. Individual projects are preferred. Small group projects are acceptable upon justification (e.g., why this is *k* number of people's worth of work). The grading of group members will be adjusted according to their contribution to the project. Course projects will involve

developing a data mining software system over the course of the semester and submitting their software, a long write-up of the project, and presenting their work in a poster session.

The grading for the course project will be split as follows (of the 40% total):

1. ***Proposal (5%):*** A 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.
2. ***Progress report (5%):*** A one-page summary of the progress, any hurdles towards timely completion of the stated objectives. 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.
3. ***Project Presentation (10%):*** Students will give a short presentation to showcase their project in class. The focus of this presentation is to demonstrate and describe what was done, report interesting observations, present key conclusions, and discuss potential limitations of the study. Students working in teams may choose to present as a group or elect one of the team members to present on their behalf. Students will not be penalized for choosing not to present individually, as long as the project itself is showcased.
4. ***Final project deliverable and report (20%):*** Students are expected to submit their project deliverable, along with a brief report. The report should include the 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.

C. Schedule

This schedule overviews the topics covered each week in class. Detail information on that week's readings and assignments will be made available on Canvas.

Week 1: Introduction to data mining (Sep 10)

- Topics covered: History, major tasks, issues, challenges, and applications of data mining; association and pattern extraction; classification; clustering; ranking; prediction; outlier detection; visualization; Item sets, matrices; sequences; time-series; streams; graphs, text; networks; multimedia; user behaviors, etc.
- Case studies: data on the World Wide Web; data in online communities; clinical data, etc.

.

Week 2: Mining item sets (Sep 17)

- Topics covered:
    - Methods: frequent pattern mining; association rules; mutual information, etc.
    - Applications: query log analysis, image classification, network analysis, etc.
- **Assignment 1 released!**

Week 3: Mining matrix data (Sept 24)

- Topics covered:
    - Methods: PCA, SVD, non-negative matrix factorization, etc.
    - Applications: recommender systems; microarray analysis, etc.

Week 4: Mining sequence data (Oct 1)

- Topics covered:
    - Methods: hidden markov models; conditional random fields; blast; etc.
    - Applications: natural language processing, biological data mining, etc

Week 5: Mining text data (Oct 8)

- Topics covered:
    - Methods: latent Dirichlet allocation, sentiment classification; etc.
    - Applications: topic modeling, social analysis, business analytics, scientific literature mining, content analysis of social media, etc.
- **Project Proposal Due!**
- **Assignment 1 Due!**

Week 6: Fall Break (Oct 15)

- Topics covered: 🍁🍂🦃🍎🏝️

Week 7: Mining stream data (Oct 22)

- Topics covered**:**
    - Methods: adaptive filtering; stream clustering; etc.
    - Applications: information filtering, social media, query log analysis, etc.
- **Assignment 2 released!**

Week 8: Mining network data (Nov 29)

- Topics covered:
    - Methods: network measures, community detection, link prediction.
    - Applications: social network analysis
- **Project Update Due!**

Week 9: Image data (Nov 5)

- Topics covered:
    - Methods: Image recognition, image classification, image-to-text generation

    - Applications: social sensing, security
- **Assignment 2 due!**

.

Week 10 Mining behavior data (Nov 12)

- Topics covered:
    - Methods:
    - Applications: advertising
- **Assignment 3 released**

Week 11: Mining time series and spatiotemporal data (Nov 19)

- Topics covered:
    - Methods: time-series analysis; outlier detection, symbolic representation; temporal mining, spatial mining, spatio-temporal mining
    - Applications: marketing, stock market prediction, etc.
- 😱**Take-home Midterm** 😎

Week 12: Map-Reduce (Nov 26)

- Topics covered:
    - Methods: map-reduce, Pig, Spark, Hive
    - Applications: massive dataset mining; log file processing
- **Assignment 3 Due!**

Week 13: Interpretable Models (Dec 3)

- Topics covered:
    - Methods: Regressions, Mixed-Effect Models, Interpretable Machine Learning
    - Applications: business intelligence, behavior analysis, causal learning

Week "14": Final Presentation Friday, December 7 -- 12pm to 3pm

- Note that this is the same week as December 3rd (Week 13)
- Final course work presented at UMSI Fall Exhibition.
- All students are expected to present their results as a poster
- Make-up presentations during scheduled Week 13 class time must be approved at least two weeks before (at instructor's discretion) unless due to exceptional circumstances

Finals Week: (Dec 17)

- **Project Report Due!** (early submission is fine too)

D. Policies

D.1. 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 assignments, homework, or the course project reports late without any effect on the overall grade. The grace period, however, cannot be used to submit the exams or quizzes 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.

D.2. Academic Conduct

D.2.1. 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.

D.2.2. 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.

D.3. 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.

D.4 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/ .

E. Suggested Readings

The readings of this course will be selected from the recent literature in major journals and conference proceedings in the field of data mining. They include, but are not limited to, the ACM KDD conference on knowledge discovery and data mining (KDD), the IEEE International Conference on Data Mining (ICDM), the ACM conference of Web Search and Data Mining (WSDM), the SIAM International Conference on Data Mining (SDM), the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Knowledge Discovery from Data (TKDD), and forums in related forums such as SIGIR, WWW, ACL, ICWSM, CIKM, etc. I also appreciate you reading the syllabus this far, so please feel free to email me a cat picture to let me you did.

E.1. Optional Textbooks

The following are *optional* textbooks that could be used for supplemental reading.

1. Han, Kamber, Pei. Data Mining: Concepts and Techniques. (Third Edition).
   This book has a focus on recent developments of techniques and applications.
2. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets
   This book focuses particularly on "big data" and distributed algorithms. Much of its content is available online for free at http://mmds.org/#book
3. Hastie, Tibshirani, Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (Second Edition).
   This book has a focus on theoretical foundations of data mining.

F. Administrative notes

1. Regular class schedule: Monday, 2:00pm to 5:00pm, starting Sep 10

Note: The classes will not *run on Michigan time but will get out 10 minutes before the scheduled end*

2. Classroom: 1255 North Quad

For directions, see https://campusinfo.umich.edu/campusmap/553 (Links to an external site.)

3. Office hours: Thursdays, 9-11am, 3341 North Quad
4. Course website: We'll be using Canvas for this course.