SI 671 - Data Mining: Methods and Applications

SYLLABUS

Class #: 174350
Instructor: David Jurgens (jurgens@umich.edu)

Meeting schedule: Wednesday, 5:30 – 8:30pm, 2185 North Quad
Office Hour: Thursdays, 4:00 – 5:00 pm, 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 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 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 format of this course will be a mixture of instruction-led tutorials and student-led presentations. The instructor will give tutorials and lead discussions in the first two weeks to introduce the basic principles and tasks of data mining. Then in every week for the rest of the semester, the instructor will start with a tutorial about the methods, followed by a student-led tutorial on applications of the methods. Depending on the background of the cohort, the instructor will decide whether to give more tutorials about the methodology of mining particular genres of data, or let students with the right background lead the tutorials.

B.1. Grading

Grading will be based on:

- One-page summaries and participation in discussion (15 summaries) (15%)
- Student-led presentation and discussion (10%)
- Mid-term exam (20%)
- In-situ data mining challenges / hands-on assignments (25%)
- Course project (30%)

B.1.1. One-page summaries (15%):

Each week, starting week 2, students are expected to write a one-page summary based on the reading assignments for the previous week. The reading assignments will cover significant papers on the topics being discussed in class. Each one-page summary is expected to discuss the key concepts described in a single paper, rather than merely stating what the paper is about. A summary of key contributions, potential limitations, suggested improvements, and ideas for future follow-up work based on the paper are encouraged. Students will be required to submit 15 total summaries and should submit at least one per week, with the exception of the week they present.

All summaries will be read by the instructor, but not graded or returned to the students. Students are expected to hand-in at least ten summaries over the semester.

B.1.2. Student-led presentation and discussion (10%):

Students are expected to lead one tutorial on the applications of the methods discussed in class based on one of the readings for the previous week. The students will lead the discussion both during and after their presentation on the topics assigned to them. In preparation of their presentation, students will be required to survey the state-of-the-art techniques from major conferences and journals for recent developments and applications of these methods and how the assigned reading relates to these developments. The discussion will focus on how to apply the methods to solve particular problems and build various applications. Each student will be in charge of one topic, depending on the size of the cohort. Typically, a presentation will last 15 minutes plus 5-10 minutes of discussion, depending on scheduling.

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. In-situ data mining challenges (25%):

Students will be required to participate in one or more in-situ data mining challenges 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 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 (30%):

A course project is required. Individual projects are preferred. Small group projects are acceptable upon justification. The grading of group members will be adjusted according to their contribution to the project.

The course project will take the format of either a software system that applies existing data mining techniques to a specific type of data, or a research experiment documented in the form of a research paper.

Examples of course projects include:

1. A de-identification tool for health records using conditional random fields
2. An efficient network clustering method for very large scale networks
3. A comparative study of community detection algorithms
4. Mining frequent sequential patterns in Twitter diffusion paths
5. A primitive study of correlating social media with the stock market
6. Author identification of historical literature (essays and novels)

The grading for the course project will be split as follows:

1. Proposal (15%): 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. Tentative deadline: Around Week 5 (early-October)

2. Progress report (10%): 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. Tentative deadline: Around Week 10 (mid-November)

3. Project Presentation (25%): 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.

**Tentative schedule:** Last lecture of the course (Thursday, December 06, 2017)

4. **Final project deliverable and report (50%):** 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.

**Tentative deadline:** Thursday of Exam week (Thursday, December 13, 2017)

C. Policies

C.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.

C.2. Academic Conduct

C.2.1. Collaboration

UMSI and the instructor 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. Read the instructions carefully and request clarification about collaboration when in doubt. Collaboration is almost always forbidden for take-home and in class exams.

C.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. You may incorporate selected excerpts 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 or MSI) student handbook available on the SI intranet for the definition of plagiarism, resources to help you avoid it, and the consequences for intentional or unintentional plagiarism.
C.3. Reasonable accommodations

The university will provide reasonable accommodations to qualified individuals with disabilities upon request. If you think you need an accommodation for a disability, please let the instructor know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information that you provide in as confidential a manner as possible. For more information, see https://ssd.umich.edu/article/americans-disabilities-act-ada (Links to an external site.) It is also the University’s policy that every reasonable effort be made to help students avoid negative academic consequences when their religious obligations conflict with academic requirements. Students who expect to miss classes, examinations, or other assignments as a consequence of their religious observance are requested to contact the instructor by the drop/add deadline. For more information see https://www.provost.umich.edu/calendar/religious_holidays.html (Links to an external site.)

D. Tentative Schedule

Please check this page periodically for a more detailed, accurate, and up-to-date schedule and reading list.

Week 1: Introduction to data mining (Sep 06)

- Topics covered:
  - History, major tasks, issues, challenges, and applications of data mining; association and pattern extraction; classification; clustering; ranking; prediction; outlier detection; visualization; etc.
- Readings:
  - [Han] – Chapter 1.
  - (Optional) Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States (*Mainly from a database point of view)

Week 2: Representations and formulations of real world data (Sep 13)

- Topics covered:
  - Item sets, matrices; sequences; time-series; streams; graphs, etc.
  - Text; networks; multimedia; user behaviors, etc.
- Case studies: data on the World Wide Web; data in online communities; clinical data, etc.

- Readings:
  - Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. [Link to an external site.]
  - (Optional) Kumar, Shamanth, Fred Morstatter and Huan Liu. Twitter Data Analytics. Springer, 2013

**Week 3: Mining item sets (Sep 20)**

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

- Readings:
  - Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan. Frequent pattern mining: current status and future directions

**Week 4: Mining matrix data (Sep 27)**

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

- Readings:
  - Jonathon Shlens. A tutorial on Principal Components Analysis
  - Yehuda Koren, Robert Bell, Chris Volinsky. Matrix factorization techniques for recommender systems
  - Karthik Devarajan. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology
  - (Optional) Spectral Methods for Dimensionality Reduction
  - (Optional) Matrix Factorization: see [Link to an external site.]
**Week 5: Mining sequence data (Oct 4)**

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

**Week 6: Mining text data (Oct 11)**

- **Topics covered:**
  - Methods: latent Dirichlet allocation, sentiment classification; etc.
  - Applications: topic modeling, scientific literature mining, content analysis of social media, etc.
- **Readings:**
  - Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter." In WWW, 2011.
  - Katherine Keith, Abram Handler, Michael Pinkham, Cara Magliozi, Joshua McDuffie, and Brendan O’Connor. "Identifying civilians killed by police with distantly supervised entity-event extraction." EMNLP 2017
Week 7: Mining stream data (Oct 18)

- Topics covered:
  - Methods: adaptive filtering; stream clustering; etc.
  - Applications: information filtering, social media, query log analysis, etc.
- Readings:

Week 8: Mining network data (Oct 25)

- Topics covered:
  - Methods: network measures, community detection, link prediction.
  - Applications: social network analysis
- Readings:

Week 9: Mining behavior data (Nov 1)

- Readings:
Week 10: Mining time series and spatiotemporal data (Nov 8)

- Topics covered:
  - Methods: time-series analysis; outlier detection, symbolic representation; temporal mining, spatial mining, spatiotemporal mining
  - Applications: marketing, stock market prediction, etc.

- Readings:
  - [Mining Time series data]
  - [Mining Spatiotemporal data]
    - Zhang, Daqing, Nan Li, Zhi-Hua Zhou, Chao Chen, Lin Sun, and Shijian Li. “iBAT: detecting anomalous taxi trajectories from GPS traces.” In Proceedings

- D Jurgens. That’s what friends are for: Inferring location in online communities based on social relationships. ICWSM 2013.

**Week 11: Deep Learning (Nov 15)**

- Topics covered
  - Methods: Convolutional Neural Nets (CNNs), Sequence to Sequence Models (seq2seq), multitask learning, graph embeddings
  - Applications: natural language processing, computer vision, user modeling
- Readings
  - A. Grover, J. Leskovec. node2vec: Scalable Feature Learning for NetworksACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.
  - Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei. LINE: Large-scale Information Network Embedding. KDD 2015.

**Week 12: Interpretable Models (Nov 29)**

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

**Week 13: Project presentations (Dec 6)**

**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.

   *This book has a focus on recent developments of techniques and applications.*
   *This book focuses particularly on “big data” and distributed algorithms. Much of its content is available online for free.*
   *This book has a focus on theoretical foundations of data mining.*
   *This book has a focus of practical applications and the use of the WEKA toolkit.*

F. Administrative notes

1. Regular class schedule: Wednesday, 5:30pm to 8:30pm, starting Sep 06
   *Note: The classes will run on Michigan time (start 10 minutes past the scheduled time).*
2. Classroom: 2185 North Quad
   For directions, see [https://campusinfo.umich.edu/campusmap/553](https://campusinfo.umich.edu/campusmap/553) (Links to an external site.)
3. Office hours: Thursdays, 4pm to 5:00pm, 1282 North Quad, starting Sep 21. By appointment until then.
4. Course website: We’ll be using Canvas for this course.