INLS 690-270 - **Data Mining: Methods and Applications**

School of Information and Library Science
University of North Carolina at Chapel Hill

Spring 2019

Course Information

Time:	Wednesday, 9:05am – 12:05pm
Room:	Manning Hall 303

Instructor:	Yue “Ray” Wang
Office:	Manning Hall 7B (Garden Level)
Office hours:	By appointment
Email:	wangyue AT email DOT unc DOT edu

Recent years have witnessed explosive growth of data generated from myriad sources, in various formats, with different quality. Analyzing information and extracting knowledge contained in these data sets become challenging for researchers and practitioners in many fields. Automatic, robust, and intelligent data mining techniques become essential tools to handle heterogeneous, noisy, unstructured, and large-scale data sets.

This is a graduate-level course focused on advanced topics in data mining. It provides an overview of recent research topics in the field of data mining. It takes a data-centered approach by surveying the state-of-the-art methods to analyze (or *mine*) different genres of data: item sets, matrices, sequences, texts, images, networks, and more.

The course will highlight the practical aspects of data mining methods and their applications, rather than theoretical aspects of statistical machine learning or optimization. The course materials will focus on how the information in different real-world problems can be formulated as particular genres, and how the basic mining tasks of each genre of data can be accomplished. To this end, the course is suitable not only for 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, information retrieval, human computer interaction, social computing, health informatics, informetrics, digital humanities, economics, and business intelligence.

Prerequisites: In this course, we will learn, use, and create computer algorithms and codes to mine data. Students are expected to have programming skills (INLS 560 or equivalence) and have taken one or more analytics-related courses (INLS 509, 512, 613, 625, or equivalence).

Learning Objectives

Throughout the course, students will gain understanding and appreciation of the fundamental concepts and a broad range of topics in the field of information retrieval. In particular, students will:

- Understand the basic principles of knowledge discovery from data;
- Understand the basic computational tasks of data mining, including pattern & association extraction, data modeling, classification, clustering, ranking, prediction, outlier detection, etc;
- Understand how information in real-world applications can be formulated and represented as different genres of data, such as item sets, matrices, sequences, time series, data streams, graphs/networks, etc;
- Demonstrate how to select appropriate data mining techniques for real-world scenarios;
- Identify 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.

Course Format

The format of this course will be a mixture of tutorials and student 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 each week in the rest of the semester, the instructor will start with a tutorial about the methods, followed by a student-led tutorial about applications. 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 to run the tutorials. Students will be required to survey the state of the art from major conferences and journals for recent developments and applications of these methods. Students giving the tutorial about applications after the instructor's tutorial about methodology are supposed to lead the discussion in the rest of the class. The discussion will focus on how to apply the methods to solve particular problems and various applications. Every student will be in charge of at least one topic, depending on the enrollment. Students who are not presenting or leading the discussion will be required to actively participate in discussion and write a one-page summary of the topic (details below).

Grading

Grade breakdown

- Active participation in class and 5 one-page reading responses: 10%

- Student presentation and leading discussion: 10%
- Take-home midterm: 10%
- Two programming assignments: 30%
- Semester-long course project: 40%

Active participation in class and 5 one-page reading responses (10%): Each week, starting Week 3, students will write a one-page response based on the reading assignments from the previous week. A total of 5 summaries should be submitted; the student have the option to choose which weeks to submit or skip. The reading assignments will cover significant papers on the topics being discussed in class. Each one-page response will follow a specific format, which will be shared on Sakai.

Student presentation and leading discussion: (10%): see "Course Format" section on previous page.

Midterm (10%): The mid-term exam will be a take-home test and will be administered around Week 11 of the semester (after Spring Break).

Programming assignments (30%): There will be two hands-on programming assignments, both closely related to the course material. In the first assignment (10%), students will warm up their data manipulation skills by building relatively simple programs to analyze data. In the second assignment (20%), students will participate in an in-class data mining challenge. The challenge will be hosted on Kaggle[1], an online data competition service. Real-world data and gold-standard judgments will be provided; students can submit and resubmit their results to the competition site and get instant feedback (evaluation metrics) from the service. Example challenge tasks include: link prediction in social networks; sentiment classification; sarcasm detection; citation prediction; community detection; book recommendation, etc.

Course project (40%): Students will apply the knowledge and skills learned in the course to accomplish a semester-long data mining project. Individual projects are encouraged. 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. 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 technical paper.

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

- ***Proposal*** (5%): A two-page proposal, describing the project topic, objectives, potential data sources, expected deliverables (software package, demo system, and/or a technical report), and a list of team members and their expected contribution to the project.
- ***Progress update*** (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.
- ***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 insights and conclusions, and discuss potential limitations of the study. Students

[1] https://www.kaggle.com

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.

- ***Final project deliverable and report*** (20%): Students are expected to submit their project deliverable (including runnable code, data, and running instructions in case of a software system), along with a report of the project. The report should include the project background, method(s) used, 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.

Grading Policy

The following grade scale will be used as a guideline (subject to any curve):

Undergraduate grading scale: A 95-100%, A- 90-94%, B+ 87-89%, B 84-86%, B- 80-83%, C+ 77-79%, C 74-76%, C- 70-73%, D+ 67-69%, D 64-66%, D- 60-63%, F 0-59%.

Graduate grading scale: H 95-100%, P 80-94%, L 60-79%, and F 0-59%.

Tentative Schedule

The following schedule is subject to change. This schedule overviews the topics covered each week in class. Detail information on that week's readings and assignments will be made available on Sakai.

Week 1, Jan. 9: **Introduction to Data Mining**

- History, major tasks, issues, challenges, and applications of data mining;
- Association and pattern extraction; classification; clustering; ranking; prediction; outlier detection; visualization.

Week 2, Jan. 16: **Representations and Formulations of Real World Data**

- Item sets, matrices; sequences; time-series; streams; graphs, etc.
- Case studies: data on the World Wide Web; data in online communities; clinical data, etc.

Week 3, Jan. 23: **Mining Item Sets**

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

Week 4, Jan. 30: **Mining Matrix Data**

- Methods: principle component analysis (PCA), singular value decomposition (SVD), non-negative matrix factorization, etc.
- Applications: recommender systems; microarray analysis, etc.
- **Assignment 1 out**

Week 5, Feb. 6: **Mining Sequence Data**

- Methods: hidden Markov models; conditional random fields; BLAST; etc.
- Applications: natural language processing, biological data mining, etc.
- **Project proposal due**

Week 6, Feb. 13: **Mining Text Data**

- Methods: latent Dirichlet allocation, sentiment classification; etc.
- Applications: topic modeling, scientific literature mining, content analysis of social media, etc.

Week 7, Feb. 20: **Mining Network Data**

- Methods: network measures, community detection, link prediction.
- Applications: social network analysis.
- **Assignment 1 due**
- **Assignment 2 out**

Week 8, Feb. 27: **Mining Image Data**

- Methods: image recognition, image classification, image-to-text generation.
- Applications: social sensing; medical diagnosis

Week 9, Mar. 6: **Mining Time Series and Spatio-Temporal Data**

- Methods: time-series analysis; outlier detection, symbolic representation; temporal mining, spatial mining, spatio-temporal mining.
- Applications: marketing, stock market prediction, etc.
- **Progress update due**

Week 10, Mar. 13: **Spring Break (no class)**

Week 11, Mar. 20: **Mining Stream Data**

- Methods: stream clustering; adaptive filtering; etc.
- Applications: information filtering, social media, query log analysis, etc.
- **Assignment 2 due**
- **Take-home midterm** (date TBD)

Week 12, Mar. 27: **Mining Behavior Data**

- Methods: generative models; correlation and causality tests
- Applications: click modeling; online advertising, etc.

Week 13, Apr. 3: **Mining *Big* Data**

- Methods: map-reduce, minhash, online learning
- Applications: log file processing.

Week 14, Apr. 10: **Mining Medical Data**

- Applications: EHR de-identification, medical concept extraction, clinical abbreviation disambiguation, drug adverse event detection, biological network analysis, consumer health vocabulary mining, etc.

Week 15, Apr. 17: **Interpretable Models**

- Methods: Regressions, rule-based models, interpretable machine learning
- Applications: business intelligence, explainable recommendation, etc.

Week 16, Apr. 24: **Project Presentations**

- Details to be announced.

Week 17, May 1: **Final Exams Week**

- **Project deliverables and report due.** Early submission is fine too.

Course Policies

Late Policy

Students should submit their assignments to the Sakai site by 11:59pm of the announced due date. Each student has *72 hours* of buffer grace period for the entire semester. If necessary, students may use it to submit any of the one-page summaries, data challenge results, or the course project report late without any effect on the overall grade. 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.* In case there is an emergency before the submission deadline, please inform the instructor as early as possible.

Collaboration

SILS strongly encourages collaboration while working on assignments, such as interpreting reading assignments as a general practice. Collaboration with other students in the course will be especially valuable in summarizing the reading materials and picking out the key concepts. However, all the work you hand in must be your own. This means that you cannot look at another student's answer and copy or re-word it as your own. Your work is a part of you; do not let someone else represent you.

If someone helps you with a homework assignment, please give them credit by writing their name(s) on the top of your homework. This will not hurt you (provided your answer is your own), but it will help them. If you are the student giving help, don't give away the answer. Rather, try to help the student arrive at the answer themselves. If you are the student asking for help, don't ask for the answer. Rather, ask about the material. It is utmost important to build up your own understanding and intuition of data mining.

Class Participation

Students are expected to read related material before every class and actively participate in discussions. Sharing your view with your peers is an important part of your education. It will sharpen your understanding of the material and help you build confidence in the area of study. Active participation in class also factors towards the 10% of your final grade.

During the semester, missing one or two classes due to legitimate reasons (e.g., travel, sickness) is fine. However, if you expect to miss more than twice during the semester, please notify the instructor one week prior to the missing class. Your attendance factors into your participation grade. If you have to miss a class, make sure to go over class material and discussions from your peers.

Laptops and Cellphones

Usage of laptop computers, cellphones, and other electronic devices is **discouraged** during class. While laptops and tablets are convenient for note-taking, they are also a source of distraction. If your laptop is open and your mind is elsewhere, it will show. As an etiquette, please mute your phone before class starts.

Honor Code

The University of North Carolina at Chapel Hill has a student-led honor system (the UNC Honor Code). We are all responsible for upholding the ideals of honor and academic integrity. The student-led honor system is responsible for adjudicating any suspected violations of the Honor Code and all suspected instances of academic dishonesty will be reported to the honor system. Information, including your responsibilities as a student is outlined in the Instrument of Student Judicial Governance. Your full participation and observance of the Honor Code is expected.

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.

Students with Disabilities

The University of North Carolina at Chapel Hill facilitates the implementation of reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability or pregnancy complications resulting in difficulties with accessing learning opportunities.

All accommodations are coordinated through the Accessibility Resources and Service Office. See the ARS Website for contact information.

Relevant policy documents as they relation to registration and accommodations determinations and the student registration form are available on the ARS website under the About ARS tab .

Recording

Please do not record the lectures in audio or video form, or share the recording on the Internet without explicit permission of the instructor.

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 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 on 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 papers in related forums such as SIGIR, WWW, ACL, ICWSM, CIKM, etc. I would appreciate you reading the syllabus this far, so please feel free to email me a note to let me know you did.

Textbooks

The following are *optional* textbooks that can be used for supplemental reading and reference.

- Jiawei Han, Jian Pei, Micheline Kamber. *Data Mining: Concepts and Techniques*. Third Edition.
[Online e-book available through UNC-Chapel Hill Libraries].
This book focuses on historical and recent developments of techniques and applications.
- Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. *Mining of Massive Datasets*.
[Much of the material is freely available online].
This book focuses particularly on "big" data and distributed algorithms..
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition.
[Freely available online].
This book focuses on theoretical foundations of data mining.
- Aston Zhang, Zachary Lipton, Mu Li, Alex Smola. *Dive into Deep Learning*.
[Freely available online].
This online book embraces a learning-by-doing philosophy: it introduces deep learning concepts and techniques with both text and executable code.