INLS 690-270 Data Mining: Methods and Applications

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

Fall 2023
(Last update: September 18, 2023)

Course Information

Time:	Mon 2:30 – 5:15 pm
Room:	Manning Hall 014
Instruction mode:	In person, on campus
Credit hours:	Three (3) credit hours
Instructor:	Yue "Ray" Wang
Office:	Manning Hall 7B (the "Garden" Level)
Office hours:	Wed 11:00 am – 12:00 pm or by appointment
	If you choose to discuss over Zoom, Please send me an email beforehand.
Email:	wangyue AT unc DOT edu

Recent years have witnessed explosive growth of data generated from myriad sources, in different formats, with varying quality. Analyzing information and extracting knowledge contained in these data sets becomes 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-centric 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 emphasize the practical aspects of data mining methods and their applications, instead of the theoretical aspects of statistical machine learning and numerical 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.

This course is designed for advanced undergraduate students and graduate students at the School of Information and Library Science. The course format is a mixture of lectures and seminars (student-led discussions).

Prerequisites: In this course, we will learn, use, and create computer algorithms and codes to mine data. Although no prerequisites are strictly required, students are strongly recommended to have already been familiar with programming (especially data science-oriented languages, such as Python or R) and basic concepts in data analytics (such as data types and schema).

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 discussion 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 expected to actively participate in discussion and write a one-page summary of the topic (details below).

Grading

Grade breakdown

- Active participation in class: 6%

- 3 one-page reading responses: 9%
- Student presentation and leading discussion: 10%
- Three programming assignments: 40% (10% + 15% + 15%)
- Semester-long course project: 35%

Active participation in class and 3 one-page reading responses (10%): Each week, starting from Week 3, students will write a one-page response based on the reading assignments from the previous week. A total of 3 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.

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

Programming assignments (40%): There will be three hands-on programming assignments, all 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 and third assignments (15% and 15%), students will develop machine learning algorithm codes to solve data mining problems. The latter two homeworks will take the form of in-class data challenges. The challenges will be hosted on Kaggle, 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[1]; sarcasm detection[2]; humor detection[3]; citation prediction[4]; community detection; book recommendation; music recommendation, etc.

Course project (35%): Students will apply the knowledge and skills learned in the course to accomplish a semester-long data mining project. Small group projects are encouraged with justification (e.g. why this is k number of people's worth of work). 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 35% 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 5-minute presentation of the project 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 presentation. Consider this as a checkpoint towards achieving the stated goals of the project. There are no penalties for changes to the proposed project, rather it may be more prudent to recalibrate or clarify the expected outcomes during this stage.
- ***Project presentation*** (15%): Students will give a 15-minute presentation to showcase their project in class. The focus of this presentation is to demonstrate and describe what was done, report interesting

[1]Previous challenge: `https://www.kaggle.com/competitions/2022fa-inls690-review-sentiment`
[2]Previous challenge: `https://www.kaggle.com/c/2019fa-inls690-sarcasm-detection`
[3]Previous challenge: `https://www.kaggle.com/c/inls690-270-funny-news-headline`
[4]Previous challenge: `https://www.kaggle.com/competitions/2022fa-inls690-citation-recommendation/`

observations, present key insights and 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.

Your presentation will be evaluated along the following dimensions:

1. Clarity: Your presentation should clearly *and* selectively communicate key concepts, facts, methods, experiments, findings, viewpoints, etc., produced from your project. Focusing a few major points and "driving them home" makes a more powerful presenation than trying to cover everything in equal detail. The purpose of the presentation is *not* to cover everything in your final deliverable. Rather, the purpose is to communicate the gist of your problem and findings in an engaging manner (next point).
2. Engagement: Your presentation should engage the audience from the start. To increase engagement, use concrete examples/cases, stories, questions, show-of-hand polls, simple games, demonstrations, multimedia, and other creative ways to keep your audience excited, entertained, actively participating in and learning from your presentation.
3. Structure: Your presentation should follow a logical and coherent structure. One common way of achieving this is to show an outline (or agenda) at the beginning of your presentation, and then remind the audience the outline when transitioning from one part to the next. This way the audience will have a "mind map" of your presentation and know what to expect in each part.
4. Uncluttered slides: Avoid using too much text on your slides. Use pictures and diagrams instead. If too much content is on one slide, try to simplify it or break it into multiple slides. If you use text on a slide, make sure the fonts are not too small for people sitting in the back. Avoid excessive use of special effects (slow fades, transitions, backgrounds, sound effects) to help your presentation focus on the substance rather than the surface.
5. Timing: In principle, your presentation (not including the Q&A part) should be properly paced such that it does not take more or less than the alloted time limit by 2 minutes. The best way of ensuring a successful presentation is to rehearse it such that you make the best use of time and deliver the best of your hard work!

- ***Final project deliverable and report*** (10%): 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 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. Detailed information on that week's readings and assignments will be made available on Canvas.

Week 1, Aug. 21: **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, Aug. 28: **Class was cancelled**

- Because of the tragic incident on UNC campus. It was fortunate that all of us in this class were safe. Let's take good care of ourselves and each other.

Week 3, Sep. 4: **Labor Day (no class)**

Week 4, Sep. 11: **Introduction continued; Real-World Data Formulation**

- 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 5, Sep. 18: **Mining Item Sets**

- Methods: frequent pattern mining; association rules; mutual information, etc.
- Applications: query log analysis, image classification, feature selection, etc.
- **Assignment 1 out**

Week 6, Sep. 25: **Well-Being Day (no class)**

Week 7, Oct. 2: **Mining Text Data**

- Methods: latent Dirichlet allocation, sentiment classification; etc.
- Applications: sentiment analysis, spam filtering, text content analysis, etc.
- **Assignment 1 due**

Week 8, Oct. 9: **Mining Matrix Data**

- Methods: principle component analysis (PCA), singular value decomposition (SVD), non-negative matrix factorization, etc.
- Applications: recommender systems; topic modeling, etc.
- Oct. 10: **Project proposal due**

Week 9, Oct. 16: **Hands-On Session: Text Data Mining Using Python**

- Practicing basic workflow of machine learning: data exploration; training/test split; feature selection and extraction; model training, evaluation, and interpretation; error analysis.
- Tools: IPython Jupyter Notebooks; numpy; pandas; scikit-learn; matplotlib/seaborn.

- **Assignment 2 out**

Week 10, Oct. 23: **Mining Image Data**

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

Week 11, Oct. 30: **Mining Sequence Data**

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

Week 12, Nov. 6: **Mining Time Series Data**

- Methods: time-series analysis; outlier detection, symbolic representation.
- Applications: marketing, stock market prediction, etc.
- **Assignment 2 due**
- **Assignment 3 out**

Week 13, Nov. 13: **Group Sharing Session**

- Sharing project progress and HW2 approaches with the class.

Week 14, Nov. 20: **Mining Network Data**

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

Week 15, Nov. 27: **Mining Medical Data**

- Methods: interpretable machine learning methods
- Applications: clinical risk prediction, clinical abbreviation disambiguation, drug adverse event detection, medical image analysis, etc.
- **Assignment 3 due**

Week 16, Dec. 4: **Project Presentations**

Week 17, **End-of-Semester Week**

- Dec. 11: **Project deliverables and report due.** Early submission is encouraged.

Course Policies

Diversity, Equity, Inclusion Statement

I value the perspectives of individuals from all backgrounds reflecting the diversity of our students. I broadly define diversity to include race, gender identity, national origin, ethnicity, religion, social class, age, sexual orientation, political background, and physical and learning ability. I strive to make this classroom an

inclusive space for all students. Please let me know if there is anything I can do to improve. I appreciate suggestions!

Mask Use (In-Person Instruction Modes)

This semester, masks continue to be encouraged yet optional in all University buildings, including classrooms. If you choose to wear a mask, we recommend choosing one that is comfortable and fits well. There are many reasons why a person may decide to continue to wear a mask, and let us respect that choice. Conversely, let us respect the choice of students who choose not to wear a mask in classrooms where it is now optional.

Late Policy

Students should submit their assignments to the Canvas site by 11:55pm 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.

University Policy: No right or privilege exists that permits a student to be absent from any class meetings, except for these University Approved Absences:

1. Authorized University activities;

2. Disability/religious observance/pregnancy, as required by law and approved by Accessibility Resources and Service and/or the Equal Opportunity and Compliance Office (EOC);

3. Significant health condition and/or personal/family emergency as approved by the Office of the Dean of Students, Gender Violence Service Coordinators, and/or the Equal Opportunity and Compliance Office (EOC).

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 .

Counseling and Psychological Services

CAPS is strongly committed to addressing the mental health needs of a diverse student body through timely access to consultation and connection to clinically appropriate services, whether for short or long-term needs. Go to their website: https://caps.unc.edu/ or visit their facilities on the third floor of the Campus Health Services building for a walk-in evaluation to learn more.

Title IX Resources

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community.

Reports can be made online to the EOC at https://eoc.unc.edu/report-an-incident/. Please contact the University's Title IX Coordinator (Elizabeth Hall, cehall@email.unc.edu), Report and Response Coordinators in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (gvsc@unc.edu; confidential) to discuss your specific needs. Additional resources are available at safe.unc.edu.

Recording

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

Usage of ChatGPT and Generative AI

ChatGPT and other generative Artificial Intelligence (AI) technologies can produce text, images, and other media. These tools can assist with brainstorming, finding information, and even reading and creating materials; however, they must be used appropriately and ethically, and you must understand their limitations. Regardless of your use of any AI tools, ***you*** are ultimately responsible for the final product of your work.

Generative AI is extremely useful; however, it has many limitations. The limitations include, but not limited to: (1) AI-generated outputs may be inaccurate or entirely fabricated even if they appear reliable or factual; (2) the sourcing and ownership of the outputs are unclear, raising ethical and intellectual property concerns; (3) the outputs are based on existing data (often scraped from online sources) and may reflect biased views and values inherited from those data.

If you decide to use generative AI in your coursework, I urge you to use it responsibly. You should critically evaluate AI-generated outputs and consider potential biases, limitations, and ethical implications when using these outputs. If you don't know whether a statement about *any item* in the output is true, then your responsibility is to research it. If you cannot verify it as factual, you should delete it. You hold full responsibility for AI-generated content as if you had produced the materials yourself. You should clearly document and report the usage of AI-generated outputs in detail if you submit material that contains them, is based on them, or is derived from them. You should use generative AI to help you think, not think for you.

Pertaining to the subject of this course, I strongly encourage you to research the mechanisms, applications, and implications of generative AI in the context of data mining and analysis!

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 or a cat picture 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.