CSC 487/587

(Advanced) Data Mining: Algorithms and Applications (4 credits)-Winter 2024

Dr. Halil Bisgin

FLINT

| MSB 119, Zoom: 95384596678 | T 8:30AM-12:10PM | Zoom Office: 2509859732 | 810 762 3374
| MR 9.00 - 10.30AM | bisgin@umich.edu | MSB 212 - (810) 762 3121 | csc587umf.slack.com |

Course Description

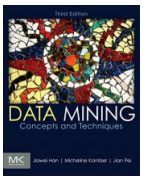
Study of the extraction of useful patterns from raw data using statistical techniques and machine learning methods. Includes use of fundamental statistical measures to summarize, pre-process and visualize raw data, and techniques for dimension reduction and data transformation. Examination of essential data mining methods, including association rules, classification and prediction, and cluster analysis, and their implementation using data mining platforms such as Weka and R

Prerequisites

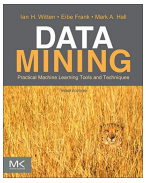
CSC 275 or CSC 276. MTH 370 recommended.

Textbook

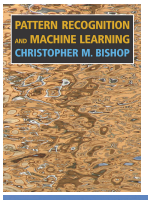
There are several textbooks, but we will follow the following textbooks closely.



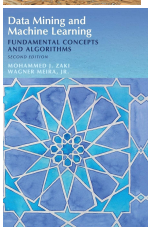
Data Mining: Concepts and Techniques, Third Edition
by Jiawei Han, Micheline Kamber and Jian Pei, 2011.
ISBN: 978-0-12-381479-1.



Data Mining: Practical Machine Learning Tools and Techniques, Third Edition
by Ian H. Witten, Eibe Frank, and Mark A. Hall, 2011.
ISBN: 978-0123748560.
(Recommended)



Pattern Recognition and Machine Learning by Christopher Bishop, Springer
ISBN 978-0-387-31073-2



Data Mining and Machine Learning: Fundamental Concepts and Algorithms 2nd Edition
by Mohammed J. Zaki (Author), Wagner Meira Jr (Author)

Tools

- Weka: `http://www.cs.waikato.ac.nz/~ml/weka/`
- R (`https://www.rstudio.com/`, `https://www.r-project.org/`)
- Python (`https://docs.continuum.io/anaconda/ide_integration`)
- Google Colab (`https://colab.research.google.com/`)
- Scikit-Learn (Machine Learning Libraries for Python) (`https://scikit-learn.org/`)
- Tensor Flow -Machine Learning Platform (`https://www.tensorflow.org/`)
- PyTorch (`https://pytorch.org/`)

Course Objectives

1. Identify data mining functionalities
2. Apply data preprocessing techniques - data cleaning, data integration and transformation, data reduction, and discretization
3. Perform attribute relevance analysis
4. Generate descriptive statistical measures in large databases
5. Mine association rules in large databases
6. Perform classification, prediction and cluster analysis.
7. Use data mining tools to perform data mining functionalities above
8. Evaluate the performance of developed models.

Tentative Schedule:

week	Topic	HW
Week 1 (Jan 9*)	What is Data Mining? What Kinds of Data Can Be Mined?	
Week 2 (Jan 16)	Introducing Data Mining Tools (R)	#1 out
Week 3 (Jan 23*)	Getting to know your data: data objects and attributes Statistical Descriptions Data Visualization Similarity and Dissimilarity Measures	
Week 4 (Jan 30)	Data Preprocessing: An overview, data cleaning data integration, data reduction	#1 in, #2 out
Week 5 (Feb 6*)	Data Transformation and Data Discretization Introducing Data Mining Tools (Weka, Python)	
Week 6 (Feb 13)	Classification: Basic Concepts, Decision Trees, K-Nearest Neighbor	#3 out
Week 7 (Feb 20*)	**Midterm Day with a shorter lecture** Classification: Bayes Classification Methods Classification: Rule Based Classification	#2 in
Week 8 (Mar 5)	Classification: Support Vector Machines Classification: Model Evaluation & Selection	mini project (graduate students only)
Week 9 (Mar 12*)	**Milestone I** (Proposal Presentations) Cluster Analysis: Introduction	#3 in, #4 out
Week 10 (Mar 19)	Cluster Analysis: Partitioning Methods Cluster Analysis: Hierarchical Methods	
Week 11 (Mar 26*)	Cluster Analysis: Density-Based Methods Cluster Analysis: Evaluation of Clustering	
Week 12 (Apr 2)	Frequent Itemset Mining Methods	#4 in, #5 out
Week 13 (Apr 9*)	Frequent Itemset Mining Methods (cont.d) Interesting Patterns, miscellanous topics	
Week 14 (Apr 16)	**Milestone II (presentations)**	#5 in

*In-person days

Important Dates	
Project Milestone I	Mar 9
Project Milestone II	Apr 16
Midterm (online)$^{\xi}$	Feb 7
Drop individual class(es) 100% refund	Jan 22, 5:00 PM.
Drop individual class(es) no refund w/ W grade	Jan 23, 5:00 PM
Withdraw(drop all classes) 50% refund w/W grade	Jan 23, 5:00 PM
Winter Break	Feb 24 - Mar 3
Final exam (online)$^{\xi}$	Tuesday, April 23 10:30 AM - 1:00 PM

ξ OASYNC students can take tests in the same evening of the announced dates.

Assessment of Student Learning Outcomes

Homework*	15%
Midterm*	15%
Project Milestone I (presentation)	20%
Project Milestone II (presentation + report)	25%
Final Exam	20%
Oral Exam	5%

*Graduate students will have additional/different questions in homework assignments and midterm.

Grading Scale

A+	97-100	C+	78-79
A	92-96	C	72-77
A-	90-91	C-	70-71
B+	88-89	D+	68-69
B	82-87	D	60-67
B-	80-81	E	0-59

Grading may be curved if the class performance warrants it. Curving can only improve grades from the grading scale above

Homework Remarks

- These are individual home works, and not group assignments. Although you are allowed and even encouraged to discuss the general concepts behind the home works and even outline of solutions with your classmates, multiple students MUST NOT work out one answer, and then submit this answer.
- I will accept late assignments. However, late assignments carry with them a 10% reduction in the grade, per week.
- All assignments must be turned in by three calendar weeks of their due date or you will receive a zero (0) grade for the assignment.
- Please note that if you are unable to turn an assignment in on the due date, then it will be considered late unless you have made arrangements with me at least twenty-four (24) hours before the assignment is due.
- Only submission via Canvas will be accepted.
- Some of the homework assignments may have to be demonstrated to the instructor.
- HW scores will be averaged based on 5 and 6 submissions for undergraduate and graduate students, respectively due to the mini-project assigned for graduate students.

Project Remarks

The project has two milestones as indicated in the schedule. Both milestones require a Power Point presentation for which contribution of each member should be clear. While Milestone I only requires to present the project proposal along with data exploration, preprocessing, and visualization, Milestone II should consist of additional content such as the summary of the approach, results, and conclusions. Furthermore, each group is supposed to submit a project report latest on **Apr 25** which should be in a conference paper* format and documents all steps performed.

The deadline to form a project group is **Jan 30**. Pleases visit Canvas→ People to self-assign yourself to a group. Failure to form your group after this date will result in 20 points loss in your 1^{st} Milestone. You are encouraged to form groups with a size of four and five members for undergraduate and graduate students, respectively. The scope of the project is such that a student by himself/herself can successfully complete the project. You must form your own groups. Talk to your peers (in class, using Slack etc) to form groups.

These are the criteria how your work will be evaluated based on your presentations. Some of them only apply for one milestone. Those are indicated in the parentheses.

1. Is the research problem introduced well? (I & II)
2. How well is the description of the data? (I & II)
3. How well is the visual design of the slides? (I & II)
4. Presented timely? (I & II)

5. Is the contribution of each member clear? (I & II)
6. Is the future work well-defined? (I)
7. Successfully applied and got results? (II)
8. Conclusions given? (II)
9. References provided? (I & II)

Your final report should be in a conference paper format as indicated above. Please visit the URL below to download the template. Following sections should be included in your paper:

1. Introduction: Introduce the problem that you want to solve. Provide some highlights about the nature of your problem and data mining techniques which would be a nice fit for your case.
2. Materials and Methods: Give a summary of your dataset. List your steps to preprocess your data and introduce your data mining techniques. Please provide a flowchart which summarizes your pipeline from data preprocessing through the implementation of the method.
3. Results: Report your results along with tables and figures. Please do not forget to put captions and refer your tables and figures within the text so that a reader can easily follow your results.
4. Conclusions: In a short paragraph, give a conclusion of your work.

* `https://www.ieee.org/conferences_events/conferences/publishing/templates.html`

Here are some resources where you can find sample data sets and research problems:

1. Kaggle repository:
   `https://www.kaggle.com/`
2. UC Irvine Machine Learning Repository:
   `http://archive.ics.uci.edu/ml/index.php`
3. National Consortium for the Study of Terrorism and Responses to Terrorism:
   `http://start.umd.edu/data-and-tools/start-datasets`

Academic Integrity

(`http://catalog.umflint.edu/content.php?catoid=2&navoid=80#Student_rights/`)

Intellectual integrity is the most fundamental value of an academic community. Students and faculty alike are expected to uphold the highest standards of honesty and integrity in their scholarship. No departure from the highest standards of intellectual integrity, whether by cheating, plagiarism, fabrication, falsification, or aiding and abetting dishonesty by another person, can be tolerated in a community of scholars. Such transgressions may result in action ranging from reduced grade or failure of a course, to expulsion from the University or revocation of degree.

It is the responsibility of all students and faculty to know the policies on academic integrity in the instructional units at the University of Michigan-Flint. Information about these policies and the appeals process is available from the appropriate administrative office of the instructional units: in the College of Arts and Sciences, the Office of the Dean of the College of Arts and Sciences; in the School of Education and Human Services, the Office of the Dean of the School of Education and Human Services; in the School of Management, the Office of the Dean of the School of Management; in the School of Health Professions and Studies, the Office of the Dean of the School of Health Professions and Studies and for graduate students, the Office of the Dean of Graduate Programs. Departments and programs within these instructional units may have specific policies and procedures which further delineate academic integrity. In such cases students are bound by the University policy on academic integrity as well as these department or program policies.

Procedural Rights of the Accused Student

A student who is charged with academic dishonesty by an instructor, administrator, or another student may be assured that he/she has the right to a fair hearing of the charges and the evidence, the right to question witnesses, to invite witnesses on his/her behalf, and to introduce whatever other evidence may be relevant to the charge.

Code of Academic Conduct

The University, like all communities, functions best when its members treat one another with honesty, fairness, respect, and trust. Therefore, an individual should realize that deception for the purpose of individual gain is an offense against the members of the community. Such dishonesty includes:

- **Plagiarism:** taking credit for someone else's work or ideas, submitting a piece of work (for example, an essay, research paper, assignment, laboratory report) which in part or in whole is not entirely the student's own work without fully and accurately attributing those same portions to their correct source.
- **Cheating:** using unauthorized notes, or study aids, or information from another student or student's paper on an examination; altering a graded work after it has been returned, then submitting the work for regrading; allowing another person to do one's work, then submitting the work under one's own name.
- **Fabrication:** fabricating data; selectively reporting or omitting conflicting data for deceptive purposes; presenting data in a piece of work when the data were not gathered in accordance with guidelines defining the appropriate methods of collecting or generating data; failing to include a substantially accurate account of the method by which the data were gathered or collected.
- **Aiding and Abetting Dishonesty:** providing material or information to another person when it should reasonably be expected that such action could result in these materials or information being used in a manner that would violate this code of academic integrity.
- **Falsification of Records and Official Documents:** altering documents affecting academic records; forging a signature of authorization or falsifying or omitting necessary information on an official academic document, election form, grade report, letter of permission, petition, or any document designed to meet or exempt a student from an established College or University academic regulation; falsification or unauthorized altering of information in any official academic computer file.
- **Identity Theft:** Assuming another person's identity or role through deception or without proper authorization. Communicating or acting under the guise, name, identification, email address, signature, or indicia of another person without proper authorization, or communicating under the rubric of an organization, entity, or unit that you do not have the authority to represent.
- **Misrepresentation and Other Acts of Academic Dishonesty:** Fraudulently obtaining and/or using academic materials that would give oneself an unfair advantage over other students or would deceive the person evaluating one's academic performance.

An attempt to commit an act prohibited by this code may be punished to the same extent as a completed violation.

Accessibility Issues

It is my intention to support the full participation of all students in the learning process of this class. Students with disabilities that may restrict their full participation in course activities are encouraged to meet with the instructor or to contact the Office of Accessibility Services.

Accommodations: The University of Michigan?Flint strives to make learning experiences as accessible as possible and complies with Section 504 of the Rehabilitation Act of 1973 and the American with Disabilities Act. The university provides individuals with disabilities reasonable accommodations to participate in educational programs, activities, and services. Students with disabilities requiring accommodations to participate in class activities or meet course requirements must self-identify with Disability and Accessibility Support Services as early as possible at (810) 762-3456 or dassflint@umich.edu. The office is located in 264 University Center, inside the CAPS Office. Once your eligibility for an accommodation has been determined you will be issued an Accommodation Letter. Please present this letter to each faculty member in each class at the beginning of the term, or at least two weeks prior to the need for the accommodation (test, project, etc.).

Available Support Services

There is a plethora of support services available to students from tutoring to mental health services. Many times students are unaware of the services available to them. One such service is tutoring:

`http://www.umflint.edu/tutoring/student-success-center-tutoring`

For other services, please check Student Success Center at

`http://www.umflint.edu/studentsuccess/student-success-center`

Generative AI Remarks

Depending on the particular assignment, you may submit material that contains AI-generated content as long as this use is properly documented to make the process transparent. For example, an assignment may require you to include an explanation as to (1) what was the original AI prompt; (2) what are some examples of incorrect data provided by the AI platform; and, (3) how did you rework and revise your final document.

The use of generative AI tools (e.g. ChatGPT, GoogleBard, Dall-e, etc.) is permitted in this course for the following activities:

- Brainstorming and refining your ideas;
- Fine tuning your research questions;
- Finding information on your topic;
- Drafting an outline to organize your thoughts; and
- Checking grammar and style.
- Further examples of acceptable AI use within a particular class are at the instructor's discretion.

The use of generative AI tools is not permitted in this course for the following activities:

- Impersonating you in classroom contexts, such as by using the tool to compose

  discussion board prompts assigned to you or content that you put into a Zoom chat.
- Completing group work that your group has assigned to you, unless it is mutually agreed upon that you may utilize the tool.
- Further restrictions on the use of AI for individual assignments are at the instructor's discretion.

Statement related to sexual misconduct

The University of Michigan-Flint is committed to preventing sexual and gender-based misconduct and offering support to those who have been harmed. Sexual assault, harassment, discrimination and all forms of sexual and gender-based misconduct have no place here. For more information, or to make a report, please visit the Equity, Civil Rights and Title IX Office (ECRT) at `https://www.umflint.edu/ecrt`.

Statement regarding alcohol and drugs

It is important for university employees and students to abide by the University of Michigan-Flint's Alcohol and Other Drugs (AOD) policy. It is your responsibility to familiarize yourself with the policy which is located here.

Frequently Asked Questions

💬 Can I submit my homework late?

💬 Yes, late homework can be accepted until one week, but there will be a 10% reduction and assignments must be turned in by one calendar week of their due date.

💬 Is that OK we ask for extension for any homework?

💬 Yes, please don't hesitate to ask. Most of the time extensions can be granted, but can't be promised.

💬 Can I pass the class if I can't do well in the tests?

💬 First, the class average and standard deviation can warrant an adjustment in calculations. Therefore, rather than comparing with the grading scale, class distribution may be more informative. Furthermore, other components of the course can contribute to major part of the final score.

Notes

I reserve the right to modify course policies, the course calendar, assignment point values, and due dates. Any extenuating circumstances that hinder your participation in the course should be discussed with me as soon as those circumstances are

known. Make-ups for graded activities may be arranged if an absence is caused by documented illness or personal emergency. A written explanation, including supporting documentation, must be submitted to me; if the explanation is acceptable, then an alternative to the graded activity will be arranged. Whenever possible, make-up arrangements must be completed prior to the scheduled activity.