CS484/584: Data Mining

Instructor: Fang-Yi Yu

Spring 2025

E-mail: fangyiyu@gmu.edu	Web: **https://cs.gmu.edu/ fangyiyu/**
Class Hours: T/Th 3:00-4:15pm	Office Hours: T/Th 4:30-5:00pm
Class Room: Innovation Hall 103	Office: Research Hall 350
GTA1: Mohammadreza Noei	Office Hours: F 10-11 am
GTA2: Roshan Dhakal	Office Hours: T 1-2 pm

Course Description

Data has been the driving force behind many of the remarkable advancements in computer science. With the exponential growth in data generation, there is a pressing need for effective techniques to extract meaningful and actionable insights from these vast collections of data. This course aims to introduce key concepts and techniques in data mining, focusing on methods such as classification, clustering, and association rule mining.

Resources

- Textbook: Pang-Ning Tan,Michael Steinbach, Anuj Karpatne and Vipin Kumar Introduction to Data Mining (Second Edition).
- Communication and class link: https://piazza.com/gmu/spring2025/cs484584
- Course website: Canvas https://canvas.gmu.edu/courses/31470

Prerequisites/Corequisites

Formally, you must have received a grade of C or better in CS 310 and STAT 344. Programming experience in Python is preferred, although Java or C will work as well (assignments will use the Python framework). Students should be familiar with probability and statistics concepts, as well as linear algebra. Please expect programming in all the assignments and class projects.

Course Objectives

1. Develop the ability to apply computing principles, along with probability and statistics, to analyze data in the field of data mining.

2. Gain a thorough understanding of programming with data mining tools and algorithms for estimation, prediction, and pattern discovery.
3. Learn to analyze problems by identifying and defining the computing requirements necessary for solutions, including data collection and preparation, functional requirements, model selection, prediction algorithms, software tools, and performance evaluation.
4. Understand and apply performance metrics in the data mining field to interpret algorithmic or model results, compare methods, and draw meaningful conclusions from data.
5. Build effective communication skills to present the steps, processes, and results of solving data mining problems to diverse audiences.

Preliminary Topics

This preliminary list of topics may change based on time constraints, the interests of the class, or

- Data and It's Various Forms
- Classification: Models, Methods and Applications
- Methods and Applications
- Fairness, Accountability, and Transparency in Data Mining and Machine Learning
- Association Rule Mining
- Anomalies and Outliers

Policies

Grading Policy

We'll calculate your final grade based on the following components. There will be no make-up or extra-credit assignments at the end of the semester; your grade should be a measure of your semester-long progress.

- Homework: 40% (best four out of five)
- Midterm 25%
- Final project 30%
- Participation 5%

Assessment

- A+ (rank >= 5%)
- A (score >= 95.0% or rank >= 10%)
- A– (score >= 90.0% or rank >= 20%)
- B+ (score >= 85.0% or rank >= 30%)
- B (score >= 80.0% or rank >= 40%)
- B– (>= 75.0%)
- C and F

Homework

There will be five homework assignments and the score of the lowest one will be dropped. Each homework assignment has a posted deadline, and late submission is not accepted unless a valid excuse is communicated to the instructor *before* the deadline. Assignments are considered individual efforts, and no sharing and discussion of problem solutions are allowed with anyone except the TAs or the instructor.

If you feel points have been incorrectly deducted, contact the grader: TA for homework and instructor for the midterm. Contesting of grades on any/all submissions must be requested within one week of receiving the grades. No grade changes will be considered after that deadline.

Midterm exam

Midterm will be on *March 20* during lecture. You are allowed one 8.5x11in sheet of notes, front and back. There will be no make up for the exam unless previously arranged (well in advance).

Final project

The final project will be graded in groups of size 3-4 people. The project will consist of the following:

1. Checkpoint, a written report that contains a research topic, a brief introduction, and a literature review of the topic,
2. Presentation in the class,
3. Final report that includes the research topic, introduction, literature, results, and discussion.

Honor Code

Please see the Office for Academic Integrity (`https://oai.gmu.edu/`) for a full description of the code and the honor committee process, and the Honor Code Policies of the Department of Computer Science (`https://cs.gmu.edu/resources/honor-code/`) regarding the course project. GMU is an Honor Code university. The principle of academic integrity is taken seriously and

violations are treated gravely. If you rely on someone else's work in an aspect of the course project, you should give full credit in the proper, accepted form. Another aspect of academic integrity is the free play of ideas. Vigorous discussion and debate are encouraged in this course, with the firm expectation that all aspects of the class will be conducted with civility and respect for differing ideas, perspectives, and traditions. When in doubt (of any kind) please ask for guidance and clarification.

Please refer to the GMU Common Course Policies provided by the Stearns Center (https://stearnscenter.gmu.edu/home/gmu-common-course-policies/), which cover any policies not directly superceded in this syllabus.

Disabilities

If you have a documented learning disability or other condition which may affect academic performance, make sure this documentation is on file with the Office of Disability Services and talk to the instructor about accommodations.