Computer Science 484: Data Mining

George Mason University

Fall 2024

Instructor: Sanmay Das
email: sanmay at gmu dot edu
Class times: Tue, Thu 3:00-4:15, David King Jr. Hall 1006
Office: ENGR 4422
Office hours: TBA and by appointment.
Questions and communication: Piazza (`https://piazza.com/gmu/fall2024/cs484002/`)
Textbook: Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar *Introduction to Data Mining (Second Edition)*. Website: `https://www-users.cs.umn.edu/~kumar001/dmbook/index.php`
Communication and Class Link: Piazza: `http://piazza.com/gmu/fall2021/cs484`
Graduate Teaching Assistants: Roshan Dhakal (rdhakal2 at gmu dot edu) and Mohammadreza Noei (snoei at gmu dot edu)
Undergraduate Teaching Assistant: Aria Asgharikamrani (aasghar5 at gmu dot edu)

1 Course Description

1.1 Overview

The amount of data available for analysis continues to increase exponentially across a broad range of areas. This leads to the need for development of techniques to discover useful and interesting information from these large collections of data. This course aims to provide an overview of key data mining methods and techniques like classification, clustering, and association rule mining. The emphasis will be on developing basic skills for modeling, prediction and performance evaluation. The course will also provide interesting applications of data mining, for example in social media analysis, text analytics, and business intelligence.

1.2 Prerequisites

Formally, you must have received a grade of C or better in CS 310 and STAT 344 (or a set of other courses deemed equivalent). 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 lots of programming in all the assignments and class projects.

1.3 Format

Class sessions will be lectures, but they could also involve in-class activities and quizzes. In addition to the textbook, other material may also be discussed, in which case pointers to appropriate reading will be provided. Grading will be based on homework assignments, the quizzes, and a project.

In this class, you are allowed to collaborate on assignments to the following extent. You are welcome to discuss problems with each other and to take your own notes during these discussions. However, you must write up solutions on your own. You must write, on the assignment, the names of students you discussed each problem with, and any external sources you used in a significant manner in solving the problem. Lack of citation of a source is a serious violation of this policy. You may not give or receive help from other students in the class on quizzes.

If you have any questions about the level of collaboration permitted, or any other aspect of this policy, please speak with the instructor or TA about it before handing in the assignment! Any deviation from this policy will be considered a violation of the GMU Honor Code.

1.4 Learning Outcomes

- The ability to apply computing principles, probability and statistics relevant to the data mining discipline to analyze data.
- A thorough understanding of model programming with data mining tools, algorithms for estimation, prediction, and pattern discovery.
- The ability to analyze a problem, identifying and defining the computing requirements appropriate to its solution: data collection and preparation, functional requirements, selection of models and prediction algorithms, software, and performance evaluation.
- The ability to understand performance metrics used in the data mining field to interpret the results of applying an algorithm or model, to compare methods and to reach conclusions about data.
- The ability to communicate effectively to an audience the steps and results followed in solving a data mining problem.

1.5 Preliminary Topics

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

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

2 Policies

2.1 Assessment and Course Grade

Your overall course score will be determined (on a curve) using the following weights. There is no absolute correspondence of scores to grades.

1. Homework assignments: 50%
2. Quizzes: 15%
3. Final project (video) presentation: 10%
4. Final project writeup: 25%

Homeworks will be submitted on a combination of Gradescope and Miner (only available on campus or through VPN) – we will make instructions available. Late assignments will not be accepted.

2.2 Make-Up Quizzes and Incompletes

Quizzes will be announced one week in advance. We will not provide make-up quizzes or incompletes.

2.3 GMU Common Course Policies

Please see `https://stearnscenter.gmu.edu/home/gmu-common-course-policies/` for important policies on academic standards, accommodations for students with disabilities, FERPA and the use of GMU email addresses for course communication, and Title IX resources and requirements that are common to all GMU courses.

2.4 Academic Integrity

Please familiarize yourself with the academic standards at Mason (`https://academicstandards.gmu.edu/`) as well as the CS Department's specific policies (`https://cs.gmu.edu/resources/honor-code/`. As stated above, collaboration in thinking through problems can be highly beneficial, and is allowed in this class. However, you may not share or look at any written material (code, answers to problems) that will be part of your or another student's submission.

Use of generative AI models: Generative AI models (including, but not exclusive to Gemini, ChatGPT, or Claude) may be used in this course as an assistant in homework assignments.

Any use must follow the fundamental principles of academic integrity and include the following statement with assignment submission: **The ideas in this submission are original and were generated by (my name). `<Generative-AI model (specify)>` was used as an editorial/coding assistant, however, I take full responsibility for the originality and accuracy of the content.**

Here are a few warnings: Large Language Models (LLMs) **hallucinate** and they do so frequently, while maintaining a tone of confidence and authority in the language they generate that would usually only be used by a human who is very confident in their answer. Recent studies have found that in technical domains they are typically wrong *at least 40% of the time*. LLMs are not human and you should be careful not to anthropomorphize them.

More generally, sharing your own original ideas with generative AI models can lead to loss of control and ownership of those ideas and coding. It can also derail your learning objectives by sacrificing the opportunity to acquire the knowledge, skills, and critical thinking taught in this course. If you rely on them, you risk being unable to perform to expectations in quizzes and other situations where they are not available. Utlimately, this could endanger your employability. This course is an opportunity to learn. Treat it as such.

2.5 Campus Closure or Emergency Class Cancelation/Adjustment Policy

If the campus closes, or if a class meeting needs to be canceled or adjusted due to weather or other concern, students should check Piazza for updates on how to continue learning and for information about any changes to events or assignments.