Computer Science 584: Data Mining

George Mason University

Spring 2022

Instructor: Sanmay Das
email: sanmay at gmu dot edu
Class times: Tue 4:30-7:10, Innovation Hall 132, and on Zoom (see Piazza)
Office: ENGR 3619
Office hours: TBA and by appointment.
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/spring2022/cs584`
Teaching Assistants: Li Zhang (lzhang18 at gmu dot edu) and Pooya Fayyazsanavi (pfayyazs at gmu dot edu)
Office hours: Li: Thursdays, 3:30-4:30 PM on Zoom (`https://gmu.zoom.us/j/91078465771?pwd=ZFBvQldjakhyL0NvdDRBaC9TUll5Zz09`)
Pooya: TBA

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. 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.

If you have already taken CS 484, you should not take this class. It will meet the CS 584 prerequisite.

1.3 Format

Class sessions will be lectures, but they may also involve in-class activities and quizzes. Quizzes will not be announced in advance, and they will all be taken online, so please bring a computer to class. 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 Course 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: 20%
3. Final project (video) presentation: 5%
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

We will not provide make-up quizzes or incompletes. We will drop your lowest quiz grade. Additional excused absences on days when there are quizzes will result in your grade being "filled in" based on performance on the ones you are present for.

2.3 Academic Integrity and GMU 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. Please make sure you are cognizant of the GMU Honor Code: `https://oai.gmu.edu/mason-honor-code/full-honor-code-document/`.

2.4 Accommodations and resources for disabilities

If you have a documented learning disability or other condition that may affect academic performance you should: 1) make sure this documentation is on file with the Office of Disability Services (SUB I, Rm. 222; 993-2474; `http://www.gmu.edu/student/drc` to determine the accommodations you need; and 2) talk with me to discuss your accommodation needs.

2.5 Safe Return to Campus Statement

All students taking courses with a face-to-face component are required to follow the university's public health and safety precautions and procedures outlined on the university Safe Return to Campus webpage (`https://www2.gmu.edu/safe-return-campus`). Similarly, all students in face-to-face and hybrid courses must also complete the Mason COVID Health Check daily. The COVID Health Check system uses a color code system and students will receive either a Green, Yellow, Red, or Blue email response. Only students who receive a green notification are permitted to attend courses with a face-to-face component. If you suspect that you are sick or have been directed to self-isolate, please quarantine or get testing. Faculty are allowed to ask you to show them that you have received a Green email and are thereby permitted to be in class. Students are

required to follow Mason's current policy about facemask-wearing. All community members are required to wear a facemask in all indoor settings, including classrooms. An appropriate facemask must cover your nose and mouth at all times in our classroom. If this policy changes, you will be informed; however, students who prefer to wear masks will always be welcome in the classroom.

2.6 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.