CS 584: Data Mining

George Mason University

Department of Computer Science

Course title and number	CS 584: Data Mining
Term	Fall 2022
Class times and location	Fri 10:30-13:10, Enterprise Hall Room 276
Piazza	piazza.com/gmu/fall2022/cs584005

1. Course Description and Prerequisites

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, including classification, regression, dimension reduction, clustering, association rule mining, recommendation, and text mining. The emphasis will be on developing basic skills for data processing, modeling, prediction, and performance evaluation. Besides, topics about dark side of data mining techniques and applications will also be introduced, like various types of unfairness, bias, filter bubble in different data mining applications.

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

2. Learning Outcomes

1) The ability to apply computing principles, probability and statistics relevant to the data mining discipline to analyze data.
2) A thorough understanding of model programming with data mining tools, algorithms for estimation, prediction, and pattern discovery.
3) 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.
4) 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.
5) The ability to communicate effectively to an audience the steps and results followed in solving a data mining problem.

The learning outcomes will be assessed based on a combination of homework assignments, exams, projects, and presentations.

3. Instructor Information

Instructor: Ziwei Zhu
Email: zzhu20 at gmu dot edu
Office: Engineering 4609
Office hours: Friday 2:30pm to 3:30pm, or by appointment
Attention: write '[CS 584]' in the subject if the email is related to this course

4. TA Information

TA: TBA
Email: TBA
Office: TBA
Office hours: TBA

5. Textbook and/or Resource Material

- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar *Introduction to Data Mining (Second Edition)*. Website: https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
- Jure Leskovec, Anand Rajaraman, and Jeff Ullman *Mining of Massive Datasets*. Website: http://www.mmds.org/

6. Grading Policies

Your final letter grade will be given based on (depending on class performance, the instructors may shift these boundaries down to raise students' grades.):

Letter grade	Points (out of 100)
A	97-100
A-	90-96
B+	86-89
B	83-85
B-	80-82
C+	76-79
C	73-75
C-	70-72
D	60-69
F	0-59

The overall course points will be determined using the following weights.

1) Five homework assignments: 40%
2) Midterm exam: 10%
3) Final exam: 20%

4) Project: 30%

Details are as follow:

Homework assignments: the first one is for students to get familiar with Python and submission systems, which accounts for 4%. And each of the rest four assignments accounts for 9%. *** Late assignments will not be accepted ***

Midterm exam: will be a one-hour closed book exam. (One cheatsheet of standard US letter size paper is allowed).

Final Exam: will be a two-hour closed book exam. (One cheatsheet of standard US letter size paper is allowed)

Project: will be a team project with 3-4 people. A team needs to submit a project proposal document (2%), a midterm pre-recorded video presentation (5%), a final report document (9%), a zip file with code and data (5%), and give a final in-class presentation (9%). Details refer to https://zziwei.github.io/CS584/project.pdf.

Attention: A missed exam cannot be made up. No late days for all assignments are allowed. So, manage your time wisely. All assignments must be performed individually.

7. Course Topics, Calendar of Activities, Project Milestones (subject to change)

Week	Date	Topic	Assignments	Project Milestones
1	08/26	Introduction to Data Mining	HW0 out	
2	09/02	Classification Basics and K Nearest Neighbor	HW0 due, HW1 out	
3	09/09	Principle Component Analysis		
4	09/16	Linear Regression		Proposal due
5	09/23	Logistic Regression and Neural Networks	HW1 due, HW2 out	
6	09/30	Deep Learning and Fairness in Machine Learning		
7	10/07	Clustering		Midterm video due
8	10/14	**Midterm Exam + project midterm presentation**	HW2 due, HW3 out	
9	10/21	Associate Rule Mining		
10	10/28	Recommender Systems: Basics		
11	11/4	Recommender Systems: Advance	HW3 due, HW4 out	
12	11/11	Fairness and Bias in Recommender Systems		

13	11/18	Text Mining		
14	11/25	Thanksgiving no class	HW4 due	
15	12/02	Project Presentation		Final report due
16	**TBA**	**Final Exam**		

8. Academic Integrity and GMU Honor Code

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/. In addition, the CS department has its own Honor Code policies (https://cs.gmu.edu/resources/honor-code/) regarding programming assignments. Any deviation from the GMU or the CS department Honor Code is considered a Honor Code violation.

9. 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 the instructor to discuss your accommodation needs.

10. 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 3 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.

11.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.