COSC 4610/5610: Data Mining

Fall 2020

Instructor:	Shion Guha	**Office Hours:**	see MS Teams
Email:	shion.guha@marquette.edu	**Office:**	N/A

Description: Techniques for extracting and evaluating patterns from large databases. Introduction to knowledge discovery process. Fundamental tasks including classification, prediction, clustering, association analysis, summarization and discrimination. Basic techniques including decision trees, neural networks, statistics, partitional clustering and hierarchical clustering.

Prerequisites: cosc 1010 (intro programming) or cosc 2100(data structures) or equivalent. I will assume that you either know or will pick up the basics of github as we will be using it heavily and it's a necessity in doing data science.

Location: Tue-Thu, 5:00-6:15 pm (hybrid online; both asynchronous and synchronous elements!)

Github: https://github.com/shionguha/cosc5610-datamining-fa20

Piazza: https://piazza.com/marquette/fall2020/cosc56104610

ProPublica data repository: https://www.propublica.org/datastore/datasets

Books: No specific books are required. I will provide links or pdfs or slides as necessary. Let's all save some money!

Objectives: This course aims to develop skills that can translate to building ethical, human centered, data products and services within organizations. At the end of the course, a successful student should be able to:

- design and develop data mining skills to understand structure and content of social data
- focus on classification models based in machine learning perspectives,
- understand a modern understanding of interpretation and inferences from classification models,
- evaluate complex, classification models from a social and ethical perspective.

Timeline:

- aug 27: introduction to the course; expectations and class policies
- sep 1-3: introduction to categorical outcomes
- sep 8-10: fitting logit models
- sep 15-17: visualizing results from logit models
- sep 22-24: interpreting results from logit models
- sep 29 - oct 1: training, testing, validation strategies
- oct 6-8: cross-validation and bootstrapping logit models
- oct 13-15: more cross-validation and bootstrapping logit models

- oct 20-22: regularization in logit models
- oct 27-29: more regularization in logit models
- nov 3-5: dealing with class imbalances
- nov 10-12: hierarchical logit models
- nov 17-19: more hierarchical logit models
- nov 24: conclusion

Grading Policy:

- Tuesdays at 11:59 pm: asynchronous weekly reading responses and discussion on piazza (50%)
- Thursdays, ongoing: synchronous class activity performance and participation (10%)
- September 24, 2020: submission: literature review and research questions (10%)
- October 15, 2020: submission: visualization and exploration (10%)
- November 5, 2020: submission: results and discussion (10%)
- November 24, 2020: submission: imbalances, outliers and other processing (10%)

This course will **not** be graded on a curve. The final grades will depend on the following scores:
A: 96 - 100; A-: 91 - 95; B+: 86 - 90; B: 81 - 85; B-: 76 - 80; C+: 71 - 75; C: 66 - 70; C-: 61 - 65; D+: 56 - 60; D: 51 - 55; D-: 46 - 50; F: 0 - 45;
There are no regrade requests.

Course Policy:

- You are responsible for your own progress. Please check your MS teams website for the course, the course github repo and piazza about course announcements and news regularly.
- This is a course that utilizes active learning principles. As a result, there are no traditional lectures and you will be required to do readings for class everyday before logging in. You will be required to post reading responses on piazza as part of class participation every **Tuesday** by **11:59 pm**.
- You are expected to ask, discuss and contribute to questions on piazza. Not only does this help you and enrich the course, but this will also count towards your grade. Please monitor piazza everyday as I will be posting readings, questions and polls there regularly.
- If you don't already, each of you will create your own free github accounts to maintain your project and any code you write that you create for this course. This is part of a data science project's lifecycle and is expected to be shown to employers by data science job applicants.
- All submission of project reports will be online via github. Any submission after the deadline will be considered late. In addition, you will be required to share any code that is part of your project submission, you will be required to send a pull request to the class github repository.
- You must cite every reference in your papers and every library (if used) in your project in ACM style. Failure to do so will be regarded as a violation of academic integrity. Please refer to the section on academic integrity for more details.
- We will follow CSCW proceedings format for final project report. CSCW is regarded as the top computational social science conference. Proceedings Formats are here. Specifically, we will be using the ACM Small Overleaf template. Submissions in any other format shall not be accepted.

- Regular attendance is essential to an active learning process. A student who incurs an excessive number of absences may be withdrawn from the class at the instructor's discretion.

Ott Memorial Writing Center: The Ott Memorial Writing Center offers free one-on-one consultations for all writers, working on any project, at any stage of the writing process. Marquette's writing center is a place for all writers who care about their writing, because every writer can benefit from conversation with an interested, knowledgeable peer. Writing center tutors can help you brainstorm ideas, revise a rough draft, or fine-tune a final draft. You can schedule a 30- or 60-minute appointment in advance (288-5542 or www.marquette.edu/writing-center), but walk-ins (in 240 Raynor or our other satellite locations) are also welcome. The Ott Memorial Writing Center also offers free workshops and hosts writing retreats.

Academic Integrity: Marquette University takes academic integrity very seriously. This is a core part of who we are as reflected by our Jesuit values. All students are required to take the Academic Integrity Tutorial. If you haven't, please go take it right now. All students are required to adhere to the Honor Pledge and follow the Honor Code. Please familiar yourself with the Academic Integrity website. There is a lot of useful information there. I take a zero tolerance policy with violations of academic integrity. All papers and projects are run through a plagiarism detection software. If you are flagged, be assured that we will have a conversation. Let's not have that conversation shall we? If I determine that you have indeed violated academic integrity, you will receive a failing grade for that component of the course or for the entire course depending on the nature and severity of the violation. Please help me to make sure that there are no such incidents.

Accessibility Policy: If you have any accessibility needs, please contact Office of Disability Services (ODS) to register them as soon as possible. ODS works with students with documented disabilities to provide accommodations for their educational needs. The course has been designed for multiple different styles of learning. However, if you have any specific learning styles that you want me to know about which would not be addressed by AES, please reach out to me within the first week of class so I can try to accommodate.