Knowledge Discovery & Data Mining

– Syllabus –

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Contact Information

- Instructor: Dr. Yong Zhuang
- You can call me Dr. Zhuang (draw on), or Yong (you own)
- Homepage: https://yong-zhuang.github.io
- E-mail: yong.zhuang@gvsu.edu
- Office: MAK D-2-234
- Office hour: Tuesday 3:00 pm - 5:00 pm, MAK D-2-234
  - Or by appointment. Send me an email to schedule a time to meet (over Zoom or in-person).
  - https://gvsu-edu.zoom.us/j/3966686420?pwd=WGxpc0N4YWcvOU9aWGxWZGYxbXZUdz09
  - Meeting ID: 396 668 6420
  - Passcode: 587684

About this course

In this course, we'll delve into the world of data mining, uncovering valuable insights from vast datasets. We will explore techniques to identify meaningful patterns, correlations, and trends, and apply them to both real-world and synthetic data. Topics covered include data preprocessing, dimensionality reduction, data visualization, predictive modeling, model evaluation, clustering, and association analysis techniques.

Course Information

- Course Homepage:
  - Blackboard: https://lms.gvsu.edu/
  - Course Website: https://gvsu-cis635.github.io
- Prerequisites
  - A solid understanding of data structures and algorithms.
  - Basic knowledge in linear algebra and statistics.
  - Proficiency in at least one programming language, coupled with practical programming experience.

Course Objectives

- Understand the fundamentals of data mining and learn basic algorithms.
- Know how to apply these algorithms effectively in real-world scenarios.
- Provide a foundational course for those interested in pursuing research in data mining.

Textbook

- Recommended: Data Mining Concepts and Techniques (4th Edition) by Jiawei Han, Jian Pei, and Hanghang Tong. Publication Date: 2023. (free at GVSU library)
- References
  - Think Python: How to Think Like a Computer Scientist by Allen B. Downey. (free)
  - Python Data Science Handbook by Jake VanderPlas. (free)
  - Applied Machine Learning in Python by Andreas C. Müller. (free)
  - Data Mining: The Textbook by Charu Aggarwal. (free)
  - Data Mining by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar.
  - Machine Learning by Tom Mitchell. (free)
  - Introduction to Machine Learning by Ethem ALPAYDIN.
  - Pattern Classification by Richard O. Duda, Peter E. Hart, David G. Stork.
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
  - Pattern Recognition and Machine Learning by Christopher M. Bishop.

Grading

- Quizzes & Homework assignments 30%
- Project 30%
- Midterm 20%
- Final Exam 20%

Grade A	Grade B	Grade C	Grades D & F
$A \geq 93\%$	$B+ \geq 87\%$	$C+ \geq 77\%$	$D+ \geq 67\%$
$A- \geq 90\%$	$B \geq 83\%$	$C \geq 73\%$	$D \geq 60\%$
	$B- \geq 80\%$	$C- \geq 70\%$	$F < 60\%$

Grading: Homework

Homework: 30%

- **Homework assignments are designed to reinforce course material and include a mix of programming tasks along with mathematical or written questions.**
- **Deadline: 11:59 pm Michigan time on the due date.**
  - *Late policy*: Assignments submitted late will incur a 10% penalty per day, capped at five days (50%). After this period, the assignment **will not be accepted**.
- **No copying or sharing of homework!**
  - But you can discuss general challenges and ideas with others

Grading: Course Project

Course project: 30%

- **Group project (3-4 people for one group)**
- **Goal: Solve a given data mining problem**
  - E.g., Analyze a dataset of historical weather patterns to predict rainfall amounts.
  - Kaggle Competition style
- **You are expected to submit a project report and your code at the end of the semester**

Academic Honesty

- Document all collaborations.
- No electronic code transfers between students.
- Code you find on the internet must be cited, with an active link to that code. That code should not solve the entirety of an assigned problem/project (i.e., don't have someone else do your project for you).
- You are encouraged to engage in conversations in online forums, but do not post solutions or solicit others to complete your work for you.
- You are encouraged to talk about problems with each other in non-technical terms (i.e., not code)
- Ultimately, you are responsible for all aspects of your submissions. You should be able to explain and defend your submission if the work is entirely your own.
  - **Suspicious cases will be reported to the Academic Honesty Committee**

Tentative Course Content

- **August 31 - September 1, 2025 Labor Day Recess:** No classes!
- **October 19-21, 2025 Fall Break:** No classes!
- **November 26-30, 2025 Thanksgiving Recess:** No classes!

Week	Topics Covered
1	Introduction to data mining, tasks, and Python basics
2	Descriptive statistics, visualization, Numpy, and Pandas
3	Data cleaning, transformation, compression, and sampling
4	Similarity and distance measures
5	Feature relationships and dependencies
6	Midterm preparation and advanced transformation
7	Midterm exam
8	Feature extraction, selection, and Markov Blanket
9	Fall Break (No Class for section 1)
10	Decision trees
11	Model evaluation, selection, and Bayesian classification
12	Regression, perceptron, clustering, and lazy learning
13	Neural networks and CNNs
14	RNNs, attention, and transformers
15	Project presentation and final exam preparation
16	Final exam

Expectations for Success

- Check Blackboard on a regular basis for announcements and assignments
- Check https://gvsu-cis635.github.io for course material.
- I'm here to help you! Come to office hours or schedule an appointment when you're feeling lost/stuck.
- Let me know when something is getting in the way of your success in this class.
- Let me know how the class and my teaching can be improved.
- Adhere to class, CIS, and GVSU policies on academic honesty.

Resources

GVSU provides opportunities for students to improve your academic skills through resources, such as

- Writing center (https://www.gvsu.edu/wc/)
- Speech lab (https://www.gvsu.edu/speechlab/)
- Research consultants (https://www.gvsu.edu/csce/)
- Library liaisons (https://www.gvsu.edu/library/about-the-university-libraries-3.html)
- Academic success center (https://www.gvsu.edu/sasc)

Warming Up

Individual introduction

- Name and what do you want to be called
- Describe your background and experience in Data Mining or Machine Learning.
- Specific topics you seek to learn from this class

About Me

- You can call me Dr. Zhuang (draw on), or Yong (you own)
- Education:
  - Ph.D., M.S. in Computer Science from the University of Massachusetts
- Experience/Interests:
  - My research interests include web application design, machine learning, and data mining.
- Personal Interests:
  - I enjoy spending time in city parks, hiking, and playing board games with friends and family.