CPSC 370: Data Mining

Adewale Sekoni

Spring, 2024

E-mail: sekoni@roanoke.edu
Office: Trexler 365B
Office Hours (in-person/zoom): MW 09:40-10:40 AM, 2:20-3:20 PM, or by appointment
Zoom: https://roanoke-edu.zoom.us/j/87142274652
Class: MWF 01:10-02:10 AM, Miller 213

Course Description

This class introduces tools used to extract information from data. In this class, we will study various statistical tools used in the description, visualization, and comparison of data. We will also learn to use machine learning algorithms to build models from our data.

Prerequisites

CPSC 170, or permission of the instructor.

Textbook (Optional)

Data mining Concepts and Techniques by Jiawei Han, Micheline Kamber, and Jian Pei.

Course Objectives

At the end of the course the successful student will be able to

- Use statistical tools to analyze large data sets.
- Present information mined from large data sets.
- Select and train machine learning models that are appropriate for the given data.
- Implement elementary machine learning algorithms.

Course Structure

Homework: On all assignments, your name must be written clearly as it appears on Inquire. Your homework must be neat and legible, you will lose points for submitting rough work.

Co-curricular Requirement: The Mathematics, Computer Science and Physics department offers a series of discussions that appeal to a broad range of interests related to these fields of study. These co-curricular sessions will engage the community to think about ongoing research, novel applications and other issues that face these disciplines. Each student is required to attend at least two of these sessions, and turn in a short paper describing the contents of the session, and your critical reflections about the topic and content. All papers must be submitted by email before April 30. List of all talks can be found here.

Reflection is an important skill to practice! Good reflection gives you an opportunity to think about what you've experienced and how it applies to your outlook. Mere restatement of events is insufficient (and, frankly, hardly worth your time). To make the task or reflection worthwhile, it should be more about you than about the event.

Here are some questions to help you write Your MCSP Co-Curricular Reflection:

1. Summarize the main ideas presented in the talk you attended.
2. How did the topics in the talk relate to math, computer science, or physics? How did these topics relate to specific courses you've taken?
3. Which topics or ideas were most confusing to you? Which topics would you like to learn more about? Elaborate on your answer! (There may be overlap in these questions)
4. How did your perspective on the main topics change after attending this talk?
5. What questions do you have about the topics presented in the talk?
6. What is something new you learned about yourself from attending this talk? Please provide a concrete example.
7. What did you learn from this talk that may help you as you move forward in your education, career, and/or life?

Grading Policy

The final grade will be computed based on the grades in the quizzes, tests, the final exam, home works and programming projects according to the following weights:

- **2%**: Co-curricular **34%**: Homework **34%**: Presentations
- **15%**: Midterm **15%**: Final exam

The final course grade will be calculated as follows:

- **[92, 100]%**: A **[90, 92)%**: A- **[86, 90)%**: B+ **[83, 86)%**: B **[80, 83)%**: B- **[76, 80)%**: C+ **[73, 76)%**: C **[70, 73)%**: C- **[66, 70)%**: D+ **[63, 66)%**: D
- **[60, 63)%**: D- **[0, 60)%**: F

Where $[a, b)$ is the set of all numbers greater than or equal to a and less than b.

Course Policies

During Class

Please do not multitask during class. I encourage you to take hand written notes as you may be allowed use them during pop quizzes.

Attendance Policy

Regular attendance in class is highly recommended. Regardless of attendance, students are responsible for all material covered or assigned in class.

Expected Number of Hours of Work per Week

You are expected to spend at least 12 hours of work each week inside and outside of class.

Office Hours

My office hours are in-person by default. I can also meet on Zoom with the link provided in the title of the syllabus.

Policies on Incomplete Grades and Late Assignments

Late assignments will be accepted for no penalty if a valid excuse is communicated to the instructor before the deadline. Otherwise, **you will receive no credit**.

Academic Integrity and Honesty

Students are expected to adhere to the Academic Integrity policies of Roanoke College. All work submitted for a grade is to be strictly the work of the student unless otherwise specified by the instructor. The policies as outlined in the Academic Integrity handbook will be enforced in the course.

Graded programs are subject to the Roanoke College Academic Integrity policies. Copying a program or a portion of a program (even a single line) or reading another person's program to obtain ideas for solving a problem is plagiarism. Other examples of integrity violation include writing code for someone else, using code written by someone else, telling someone else how to solve a problem or having someone tell you how to solve a problem (and using his/her method). These cases apply to any work that is handed in for a grade under the instructor's assumption that the work is your own. Unless specified otherwise by the instructor, discussion among students should be limited to general discussion of concepts and language details, not specific aspects of a solution to the assigned problem

Students are allowed to incorporate generative AI tools such as ChatGPT to enhance the quality of their assignments. It is essential, however, that you provide the base solution and understand the enhancements provided by the AI. You cannot use an AI to provide a solution you can not competently defend. In the context of papers, AI can be employed to critique and offer suggestions for changes in sentence structure. Whenever significant AI assistance is utilized, it is mandatory to acknowledge it in your work and provide a brief description of the specific ways in which it was applied.

Topics

- 8.3 Bayes Classification Methods 350
- 8.4 Rule-Based Classification 355
- 8.5 Model Evaluation and Selection 364
- 8.6 Techniques to Improve Classification Accuracy 377

6. Chapter 9 Classification: Advanced Methods 393

- 9.1 Bayesian Belief Networks 393
- 9.2 Classification by Backpropagation 398
- 9.3 Support Vector Machines 408
- 9.4 Classification Using Frequent Patterns 415
- 9.5 Lazy Learners (or Learning from Your Neighbors) 422
- 9.6 Other Classification Methods 426

7. Chapter 10 Cluster Analysis: Basic Concepts and Methods 443

- 10.1 Cluster Analysis 444
- 10.2 Partitioning Methods 451
- 10.3 Hierarchical Methods 457
- 10.4 Density-Based Methods 471
- 10.5 Grid-Based Methods 479
- 10.6 Evaluation of Clustering 483

Exams

Midterm: Friday, March 1st

Final: 8:30-11:30 AM, Saturday, Apr 27th