CPSC 370: Data Mining

Adewale Sekoni

Spring, 2023

E-mail: sekoni@roanoke.edu
Office: Trexler 365B
Office Hours (in-person/zoom) : MWF 10:50-12:20 PM, or by appointment
Zoom: https://roanoke-edu.zoom.us/u/k0QsdL4gt
Class: MWF 09:40-10:40 AM, Trexler 363

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 use machine learning algorithms to build models from our data.

Prerequisites

CPSC 170, or permission of the instructor. Familiarity with Unix is assumed.

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 before May 15**. List of all talks can be found here.

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%**: A **90-92%**: A- **86-89%**: B+ **83-85%**: B **80-82%**: B- **76-79%**: C+
- **73-75%**: C **70-72%**: C- **66-69%**: D+ **63-65%**: D **60-62%**: D- **< 60%**: F

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 on Zoom by default. The link can be found on the course Inquire page. I will be glad to come into the office for an in-person meeting if you make an appointment at least a day before, otherwise, I may not be able to make it.

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

Topics

1. Chapter 1 Introduction 1
   - 1.1 Why Data Mining? 1
   - 1.2 What Is Data Mining? 5
   - 1.3 What Kinds of Data Can Be Mined? 8
   - 1.4 What Kinds of Patterns Can Be Mined? 15
   - 1.5 Which Technologies Are Used? 23
   - 1.6 Which Kinds of Applications Are Targeted? 27
2. Chapter 2 Getting to Know Your Data 39
   - 2.1 Data Objects and Attribute Types 40
   - 2.2 Basic Statistical Descriptions of Data 44
   - 2.3 Data Visualization 56
   - 2.4 Measuring Data Similarity and Dissimilarity 65
3. Chapter 3 Data Preprocessing 83
   - 3.1 Data Preprocessing: An Overview 84
   - 3.2 Data Cleaning 88
   - 3.3 Data Integration 93
   - 3.4 Data Reduction 99
4. Chapter 6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods 243
   - 6.1 Basic Concepts 243
   - 6.2 Frequent Itemset Mining Methods 248
   - 6.3 Which Patterns Are Interesting?—Pattern Evaluation Methods 264
5. Chapter 8 Classification: Basic Concepts 327
   - 8.1 Basic Concepts 327
   - 8.2 Decision Tree Induction 330
   - 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 and Break Days

Midterm: 09:40-10:40, Friday, March 3rd

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