CPSC 370: Data Mining

Adewale Sekoni

Spring, 2020

E-mail: sekoni@roanoke.edu
Office Hours: MWF 12:00-2:00pm, or appointment
Office: Trexler 365B

Class Hours: MWF 9:40-10:40pm
Classroom: Trexler 363

Course Description

This class Introduces students to the tools used in the extraction of information from data. We study of various statistical tools and machine learning algorithms. We will also cover the preparation of data sets for use with machine learning algorithms.

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 (PDF available online).

Course Objectives

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

- Use statistical tools to analyze large datasets.
- Select and implement machine learning techniques that are suit-able for the applications under consideration.
- Recognize and implement various ways of selecting model parameters for different machine learning techniques.

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 his/her critical reflections about the topic and content. **These papers are due in class within a week of the session**. A paper submitted beyond a week from the event being discussed in the paper will NOT be accepted. The MCSP Conversation Series website has the schedule of talks in the series.

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 **54%**: Homework
- **20%**: Midterm **24%**: 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

If you use an electronic device such as a tablet or a laptop for notetaking or to read the textbook, the content that is open on the screen should be strictly restricted to documents and pages of relevance to the class. For example, you should not have any social media websites open in your browser window, even if it is in a tab that is not currently in focus. I encourage you to take handwritten notes as you may be allowed use them during pop quizzes. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.

Attendance Policy

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

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

- What is data mining?
- What kinds of data can be mined?
- Knowing your data.
- Descriptive Statistics.
- Data visualization.
- Measuring data similarity and dissimilarity.
- Data preprocessing.
- Data scrapping.
- Data cleaning.
- Data Reduction.
- Histograms.
- Clustering.
- Inferential Statistics.
- Decisions Trees.
- Bayes Classifications.
- Rule based Classification.
- Model evaluation.
- Support Vector machines.
- K-nearest neighbors classifiers.
- Artificial neural networks.
- Clustering

Exams

Midterm: 9:40-10:40, Monday, February 24th

Final: 2:00-5:00pm, Thursday, April 23rd