Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining - Final Project

Due: December 12, 2023(100 points)

NAME(S)

PROJECT GIT REPOSITORY

1 Introduction

Introduce us to your problem, why it's important, and what motivated you to take it on.

2 Background

This section should describe how this problem arises, who is affected, and any literature that you plan on reviewing or survey. These references should relate to both the applications of the work as well the technical approaches to be taken. Emphasize why they are appropriate and relevant.

If you're choosing a project besides Greenstand, make sure you describe the specifics of your data and what information that you believe would be most useful from it. If you're using the Greenstand data, articulate what features and data you plan on leveraging.

3 Approach

This section should provide the meat of your work. I expect everyone on the project to contribute to this section. Teammates may try several different approaches, but please aggregate the information across the approaches rather than individually presenting them. For example, if person A tries random forests, person B tries deep neural networks, and person C tries naïve Bayes, present these on the same graph, and talk about them in aggregate. (Do *not* have three subsections devoted to each approach.)

Your work is evaluated on the criteria relating to the presentation of the material and the technical approach. Please review the subsections that will serve as a grading rubric.

3.1 Data and Data Analysis

Amount of Data - In this course, we are mining data, which means that there should be enough rows / samples for the mining to be interesting and the conclusions to be statistically significant. This number varies with number of features and labels, but 100,000 rows should be sufficient. It must not be able to be intuited by hand, at least easily.

Data Analysis and Exploration - Please analyze your data before doing any modeling. Understand the distributions of your data. Determine feature correlation with the label (if you have labels) and each other.

3.2 Implementation

Correctness Ensure that your algorithm works and appropriately models the phenomenon without overfitting (or underfitting). For example, if you have class imbalance of $P(A) = 0.9$ and $P(B) = 0.1$, and your implementation should not claim 90% accuracy by simply always prediction A.

4 Results and Evaluation

Well-Reasoned Evaluation Clearly define the evaluation metrics you're using and why. Properly evaluate and draw conclusions. Determine the confidence you have in the outcomes of your processing and modeling.

5 Conclusions

Discuss what you've learned. How do you think it can be applied? If you had more time, what are the future directions? If there's literature on future directions, include them.