CISC 5790: Data Mining
Fordham University, Spring 2022 Prof. Yijun Zhao

Course Project

Due: May 9

Now that you are all experts in data ming, the time has come to explore an interesting project of your choice. You can work in a group of no more than 3 people. Please finalize the process of finding teammates ASAP and sign up your team following the link given below.

Each team has to create all deliverables from scratch. In particular, it is not allowed to copy another team's code or text and modify it. If you use publicly available code or text, you need to **cite the source** in your report!

Project key events:

- Team sign ups are due on **2/28** in HW2.
  `https://docs.google.com/spreadsheets/d/1xqOHBx-uCCkQ216RgQF-36dizmZuyzuE-xTGTR_LjBU/edit#gid=0`
- Final code and project report are due on **5/9**.
- Each team present their project in our last class session (before the final exam).

The project will be graded in a scale of 100, which breaks down as follows:

- Program (code and execution results match the report) 50%
- Report 50%

1 Project Details

You are free to choose your own project, but you need to make sure you are working on a course-related task. Here is the default project to use if you do not have your own preference.

1.1 Introduction

This project requires you to design and implement a model, run experiments on a real-world dataset, and write a report explaining your experimental results. Specifically, you will build binary classifier(s) that can interpret the data format specified below and report interesting performance measures such as accuracy, recall, specificity, precision, etc.

1.2 Algorithm

Your model should be based on the algorithms learned during this course. Your model can be a combination of methods and should incorporate one or more data mining techniques when the situation arises. These techniques may include (and are not limited to):

- Different treatment of various types of features: continuous, discrete, categorical, etc.
- Proper imputation methods for missing values
- Handling imbalanced dataset

1.3 Data

You'll be examining the behavior of your classification algorithm on a dataset from the UCI machine learning lab. The dataset is represented in a standard format, consisting of 3 files. The first file, `census-income.names`, describes the categories and features of the dataset. It also has some empirical results for your reference. The other two files are `census-income.data` and `census-income.test`, containing the actual data instances, formatted at one instance per line, as follows:

$F_1^1, F_1^2, \ldots, F_1^k$, label_1

$F_2^1, F_2^2, \ldots, F_2^k$, label_2

$\vdots$

$F_n^1, F_n^2, \ldots, F_n^k$, label$_n$

where F_i^j, label$_i$ $(i = 1, \ldots, n, j = 1, \ldots, k)$ represent the value of the j^{th} feature and class category for the i^{th} instance respectively.

This dataset was extracted from the census bureau database. Each instance contains an individual's educational, demographic, and family information. The prediction task is to determine whether a person makes over 50K a year. You should use `census-income.data` to train your classifier and use `census-income.test` to evaluate the performance of your learning algorithm.

2 Your Mission...

Deliverables for this project are:

- Code to implement your model for the task(s) you have chosen.
- **A README file, with simple, clear instructions on how to compile, train, and test your code**.
- Dataset(s) used for the model. If too large, you should provide your model and sample test data.
- A discussion of data mining algorithms and techniques employed in your approach.
- A report analyzing your model's performance and discoveries. At a minimum, you should provide training and test evaluations.

3 How to turn in your code

- Zip all your files (code, README, written report, etc.) in a zip file named $\{firstname\}\_\{lastname\}\_CS5790\_project.zip$ and upload it to Blackboard.
- **Only one person in your group needs to turn in the code and the report. Make sure every team member's name is listed on the cover page of the report.**