CS6220 Data Mining Techniques – Spring 2015
Course Project – due: 4/30/15

1 Introduction

This project requires you to implement a classificatiion algorithm, run experiments on a real world dataset, and write a report explaining your experimental results. Your program should be a two-class classifier that can handle datasets with an arbitrary number of features and instances. The language of implementation is up to you — the only requirement is that your program be able to interpret the data format specified below, and be able to classify instances and produce interesting statistics such as accuracy, variance, ... etc. You are free to construct whatever user interface for your program, but you must *fully document* your interface.

2 Algorithm

- Your algorithm should be based on one of the classification algorithms learned during the course. Usually a straight forward implementation of one method will not lead to satisfactory performance. Your algorithm can be a combination of methods and should incorporate one or more data mining techniques when the situation arises. These techniques include (and certainly not limited to):
  - Different treatment of various type of features: continuous, discrete, categorical, etc.
  - Proper imputation methods for missing values
  - Handling imbalanced dataset

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 emprical 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 repectively.

The data you will be examining was extracted from the census bureau database. Each instance contains an individual's educational, demographic and family information. 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.

4 Your Mission...

Deliverables for this project are:

- Code to implement the classification algorithm for the data file formats given above
- **A README file, with simple, clear instructions on how to compile and run your code**
- Testing statistics for the application of your learning algorithm. At a minimum you should provide training set accuracy, test set accuracy
- A discussion of data mining techniques employed in your algorithm
- A report analyzing the behavior of your algorithm on the dataset, including any unusual or anomalous (in your opinion) behavior

5 How to turn in your code

- **Your program must run on CCIS machines in WVH 166 lab**
- **Zip all your files (code, README, written report, etc.) in a zip file named** $\{firstname\}\_\{lastname\}\_CS6220\_project.zip$ **and upload it to Blackboard**