CS6220 Data Mining Techniques – Fall 2016 Sec01 **Instructor: Yijun Zhao**

Course Project – due 12/12/16

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. 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 model should be based on the algorithms learned during the course. Usually a straight forward implementation of one method will not lead to satisfactory performance. Your model 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 model on a dataset from the UCI machine learning lab. The data you will be working on is a real world dataset, representing 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. The data contains 50+ attributes such as race, gender, age, HbA1c test result, diabetic medications, etc. **Your task is to predict, as accurately as possible, whether a person will be readmitted within 30 days after being discharged from the hospital**. A more detailed description of the dataset can be found under

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

Following the above link, you can download the dataset under the tab `Data Folder` in the zipfile `dataset_diabetes.zip`. The zipfile consists of two files. The first file, `IDs_mapping.csv`, describes the categorical codes used for some attributes. The second file, `diabetic_data.csv` contains 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.

Note:

- **The dataset has 3 class labels `'NO'`, `'>30'`, `'<30'`. To simplify the project, we will treat `'NO'`, `'>30'` as the same class. That is, you will pre-process data file such that your project is a binary classification problem.**
- **After the pre-processing, your program should output the following for a sanity check:**

  total number of class 0 (`'NO'` and `'>30'`) instances: xx
  total number of class 1 (`'< 30'`) instances: yy

- **You must split the data using the k-fold cross-validation technique to train and evaluate the performance of your learning algorithm.**

4 Your Mission...

Deliverables for this project are:

- Code to implement your model for the data file formats given above.
- **A README file, with simple, clear instructions on how to compile and run your code**.
- Statistics of your model. At a minimum you should provide training and test accuracies.
- A discussion of data mining algorithms and techniques employed in your model.
- A report analyzing your model's performance and discoveries on the dataset.

5 How to turn in your code

- **Your program must run on CCIS machines in WVH 102 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.**