COSC6339: Big Data Analytics
Machine Learning with a Summarization Matrix

1 Introduction

In this project, you will use compute a correlation matrix, a PCA model, followed by a classifier. Your project will need to be able to generate a model based on a training data set and score (compute the prediction) on an independent test data set with the model. You will use some tecnique to reduce the numbner of dimensions/variables.

2 Machine Learning Models

The input will be a data set X with n vectors (points) having d numeric dimensions and one binary class attribute Y.

You will have two choices to reduce d: (1) variable selection, via the correlation matrix, computing correlations between each independent variable and Y. (2) PCA to reduce d and get representative dimensions in the PCs.

Variable selection: Compute correlation matrix ρ based on Γ. using ρ, select variables highly correlated to Y.

PCA for dimensionality reduction: Compute correlation matrix ρ based on Γ. Compute PCA on ρ, the correlation matrix of the data set. You will identify a small subset of dimensions that are as independent as possible. You will use PCA taking the top PCs and identifying representative attributes in each PC (stronger correlations to PC). Compare speed with existing PCA in an existing math library.

Classification: the classification models you must program are below. Teams 1,4,7,.. will choose Model 1. Teams 2,5,8,.. will choose Model 2. Teams 3,6,9,.. will choose model 3.

1. Naive Bayes using Γ
2. Logistic regression predicting Y using Gradient Descent (or Ascent)
3. Class decomposition using Γ (K-means on each class)

You must compare accuracy and speed with some existing classification model. Choose one that you like; no need to compare against all of them.

1. Decision tree (any).
2. Forest of decision trees (boosted).
3. SVM (Gaussian kernel).
4. Discrete Naive Bayes (histograms).
5. Standard logistic regression (iterative, slow).

For the classification problem given the input data set, you will need to build a training dataset and a test dataset. You will need to generate a model (classifier/regression) using the training data set. Then, you will apply this model to the test data set and produce a new data set with the predicted class. You will need to generate a Type I, Type II error table (true/false positives, true/false negatives).

3 Programming

- Programming language: Python.
- Optional: using C++ to accelerate a demanding computation.
- System: you can get started in your laptop, but you must upload and test in our Linux server by the deadline (login instructions to be posted later).

Program call: your program must run from the command line, with a list of parameters specifying task, model, input data set, train/test. You have flexibility in chooing parameters.

Examples:

```
python3 dimredclass.py "task=dimreduction;model=pca;input=med1.csv".

python3 dimredclass.py "task=classify;model=nb;input=med1.csv".
```

4 Input Data Sets

There will be two biomedical data sets: choose one. It is expected each data set will be used by 50% of teams.

5 Deliverables and Demonstration

You will need to submit your programs in the Linux server and a technical report as specified below. Your program must run from the command line with all input parameters. There should be a README file on how to run the program. The TAs will arrange a session to give a 10-minute presentation and demonstration (outside class time).

- PCA analysis: model, input, output, PCs, loadings, selected attributes.
- classification: model, input, classification accuracy, best attributes.
- Comparison: compare accuracy and speed with built-in models in existing libraries.