CS 557 STATISTICAL PATTERN RECOGNITION AND LEARNING
FALL 2016
ASSIGNMENT 1

DUE: 3rd September, 2016.

PROBLEM

Background Reading: Bishop chapter 2.

1. Read the dataset trainNaive.txt and its corresponding label file trainNaiveLabels.txt for the training data and its corresponding labels. This file's corresponding .csv file is placed on Kaggle according to the required Kaggle's format of having the first row as the header and first column with data ids.

2. Generate your naïve Bayes' model using the training set

3. Determine the predictions on the test data given in testNaive.txt file

4. Generate the .csv format for Kaggle and place your test set predictions in it. A sample solution file 'naiveSolutionLabels.csv' is there to help you look at the format. You can also use the Matlab script writeKaggle to generate it. The sample submission file has random predictions. The URL for Kaggle's competition is:

https://inclass.kaggle.com/c/sprfast-naive-algorithm

7. Report the training set error as balanced error rate (BER). Also report the accuracy (AUC) that you get from Kaggle's website on the test set.

BER = 1/2(totalErrorsOfClass+1/totalOfClass+1 + totalErrorsofClass-1/totalOfClass-1)

You can also get the AUC on the training set by using Matlab's function perfcurve

8. Once you have built a model, suggest a way to improve the performance by looking at the probabilities of different features. Is it possible to drop a few features and improve the classification performance? On what basis would you drop these features?

HINTS FOR CODE

NOTE: Be systematic when implementing your program. You can implement the following functions along with a main script in Matlab for the above steps

```
trainX = load('...','-ascii');  %replace ... with filename
%above is a built in function in Matlab for reading text files

[probVecClass0 probVecClass1 prior0 prior1] =  learnProb(trainX,trainLabels)
%this function should take the data matrix trainX and the corresponding labels as input parameter and return the corresponding parameters for building your classifiers

predictedLabels = testMAP(X,probVectorClass0,
                                        probVectorClass1,prior0,prior1)
predictedLabels = testML(X,probVectorClass0,
                                        probVectorClass1,prior0,prior1)
%The above two functions are for making predictions using MAP or ML in naïve Bayes'
```

Once you have implemented the above functions write a main script that:

a. Reads training data
b. Finds the model parameters (relevant to the distribution to use)
c. Reads the test data and classifies the test data

Note: You can use matlab's helper functions like load, sum, mean, cov, plot etc. but NOT the Bernoulli distribution functions/naive Bayes' functions provided by it.

TO SUBMIT

1. Make a folder with your roll number as folder name. Put Matlab's source code in it and place it in the 'submit assign1' folder on xeon. PLEASE DO NOT EMAIL
2. **Hard copy** of a report which is **not more than two pages** long that describes your experiments and your observations. Make sure you make a table of results to summarize your results. It should have the following:
    a. training set and test set accuracy/error rate for MAP and ML
    b. Your suggestion on improving the result as given in step 8