**CS 557 STATISTICAL PATTERN RECOGNITION AND LEARNING**

**FALL 2015**

**ASSIGNMENT 1**

**DUE: Wednesday, 2nd September, 2015.**

**PROBLEM**

**Background Reading:** Bishop chapter 2 andAlpaydin’s chapter 5.

1. Read the dataset lucapTrain.txt and its corresponding target values. The test data points and corresponding targets are in test files.

2. This is Lucap lung cancer data for lung cancer vs. no lung cancer used in the causality challenge. Please see <http://www.causality.inf.ethz.ch/data/LUCAS.html> for more information.

3. Determine the parameters of the training model when multivariate Bernoulli distribution is fitted to the data and MAP (maximum aposteriori) estimator or ML (maximum likelihood) estimator is used. YOU HAVE TO USE TRAIN FILE. **Important:** You have to use Laplacian smoothing to get smoothed probability estimates. This will also ensure that your model does not collapse if there are any constants in the data.

4. Classify all points in the training and test files using MAP and ML classifier based on multivariate Bernoulli distribution. **Important**: The parameters should be computed from training file.

7. Report all error rates as balanced error rates. It is computed using the average of error rate of class (+1) and class (-1) as following. You can also use the **balancedErrorRate** MATLAB function provided with this assignment:

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

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

[probVectorClassPlus probVectorClassMinus priorPlus priorMinus] = learnBernoulli(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 = testBernoulliMAP(X,probVectorClassPlus,

probVectorClassMinus,priorPlus,priorMinus)

predictedLabels = testBernoulliML(X,probVectorClassPlus,

probVectorClassMinus,priorPlus,priorMinus)

%The above two functions are for making predictions using MAP or ML by assuming that the distribution of data is a Bernoulli distributionf

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

1. Reads training data
2. Finds the model parameters (relevant to the distribution to use)
3. 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 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:
   1. training set and test set accuracy/error rate for MAP and ML using Bernoulli distribution