**1601 Machine Learning (EE 514/CS 535)**

**Fall 2016**

**Assignment 2**

**Deadline: Friday 7th Oct, 2016**

**Time: 11:55 pm**

**Q1.)** **Classify the data set using Naïve Bayes.**

The task involves the implementation of the classifier to an interesting problem of speech classification for Parkinson’s disease.

1. Data has been divided into 80/20 ratio of training and test data set.
2. Evaluate the performance of the classifier in the form of ROC curve
3. Provide a general intuition based on your results indicating the qualitative data analysis.
4. What will happen to the model if size of the training set is reduced?

**Dataset description:**

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voices recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

**Q2.)** **Perform classification by estimating the MLE estimates of multivariate Gaussian distributions**. The dataset for this task is the Banknote authentication dataset available at [https://archive.ics.uci.edu/ml/datasets/banknote+authentication](https://archive.ics.uci.edu/ml/datasets/banknote+authentication%20) . For this task we are providing you with the dataset that is already split into training (80%) and test datasets (20%).First four columns represent features: Variance, skewness, Kurtosis and Entropy, whereas the last column represents the class of bank notes: 0 = Fake, 1 = Genuine.

1. For each class learn MLE estimates of a multivariate Gaussian distribution i.e. mean vector and covariance matrix.
2. Classify the test data by determining that class for each record for which  is greater. Report your result as a matrix where the first column represents the original classes from test data and the second column represents predicted classes.
3. Evaluate the performance of the classifier in the form of ROC curve

Determine the accuracy of each classifier as number of correct predictions on test set divided by the number of examples of that class in test set.

**Instructions:**

1. **No late submission will be accepted.**
2. **Use Matlab for this assignment.**
3. **Use LaTex for writing your analysis Report. LaTex file along with pdf should be submitted (BONUS). Otherwise you can submit a word doc as well. You don’t have to write minor details, report shouldn’t exceed 2 pages, however do write how you compute the estimates of the parameters.**
4. **Your code for each task should generate an output.txt file that has a column of predictions for each data point in test data. PLEASE the order of the output should match the order of data points in test data.**
5. **Don’t forget to include your output.txt file for each task in separate question folder.**
6. **Your submission folder should be named on your roll number.**
7. **So here is how your assignment folder structure should be:**

Submitted folder:

15030045

Naïve Bayes folder

Multivariate Gaussian Folder

Report File

**Matlab code file Matlab code file**

**Output.txt file Output.txt file**