CS 557: STATISTICAL PATTERN RECOGNITION AND LEARNING
FALL 2016
ASSIGNMENT 2
DUE: Wednesday, 28th September, 2016.

PROBLEM

Background Reading: Bishop chapter 2 or Theodoridis & Kourtoumbas chapter 2

Data

- This data is derived from UCI machine learning repository:
  Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
- It is the same dataset with noise added to it

Kaggle URL
https://inclass.kaggle.com/c/one-more-iris-competition

1. Read the training file
2. Determine the parameters of the training model when multivariate Gaussian distribution is fitted to the data and MAP (maximum aposteriori) estimator or ML (maximum likelihood) estimator is used. You have to try out the following:
   a. Identity covariance matrix for the three classes
   b. Diagonal covariance matrices which are same for three classes. So take only the diagonal entries of the covariance matrix of the entire data.
   c. Same covariance matrix for all classes. Take the covariance matrix of the entire data and use it for all classes
   d. Take different diagonal covariance matrices for all classes
   e. Take different covariance matrices for all classes.
   f. Classify using both MAP and ML for all cases (above a-e) for both training and test sets. So you have to report a total of 20 results. Make a nice table to illustrate your results. The test set predictions have to be submitted on kaggle.

7. Report all results as accuracy rate, i.e., number of correctly classified/total examples.

8. Suggest a method for selecting features to improve the accuracy of the classifier.

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

%function to learn Gaussian distribution parameters when given training data and corresponding labels
[meanClass1 meanClass2 meanClass 3 covMatrixClass1 covMatrixClass2
covMatrixClass3 prior] = learnGaussianParameters(trainX,trainLabels)

predictedLabels = testGaussianMAP(X,…) %self explanatory. Give the parameters this requires
predictedLabels = testGaussianML(X,…) %self explanatory. Give the parameters this requires

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 Gaussian 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 assign2' 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 rates for MAP and ML using the different covariance matrices mentioned earlier
    b. training set and test set accuracy/error rate for MAP and ML using Gaussian distribution for all the cases mentioned earlier
    c. Your comments/conclusion in the report are mandatory. Write a paragraph about your experience with this classifier and you can compare the ease of implementation etc. with the classifier for assignment 1.
    d. Your suggestion for feature selection relevant to this classifier and your results if you practically implemented it.