ECE 532
Homework 6

Due Thursday February 24 at the beginning of class

Consider the Iris flower dataset `iris.mat`. The dataset contains features for three different classes of flowers:

0	–	Iris Setosa
1	–	Iris Versicolour
2	–	Iris Virginica

There are 50 examples for each class in the dataset. Each example has a four dimensional feature vector:

1	–	sepal length in cm
2	–	sepal width in cm
3	–	petal length in cm
4	–	petal width in cm

These data are organized in the file `iris.mat` in the following format. Each row corresponds to an example. Columns 1,2,3 4 are features for each example, and column 5 is the corresponding class label 0,1,2.

1. Split the training data for each class into two subsets of sizes $n = 25$ and $m = 25$. Use the first $n = 25$ examples to maximum likelihood estimates of the mean vector and covariance vector for each class. Then use the corresponding multivariate Gaussian densities with these MLEs to classify the remaining $3m = 75$ examples. Report the sample means and covariances for each class and the error performance of the trained classifier. Specifically, construct a 3×3 table of the outcomes of the classifier with entry (i, j) corresponding to the number of times a feature with true label j was classified as i.

2. How does the classifier perform as the number of training data n is varied? Specifically, construct a classifier using $n = 15, 20, 25, 30, 35, 40$ and test it on the remaining $3m$ examples in each case. Plot the total probability of error as a function of n.

3. Consider a simple dimensional reduction based on discarding one of the four features. Which feature would you discard and why? Repeat the error analysis above in 2 in this reduced feature space. Plot the total error as a function of n again, and compare to the previous results using all four features.