Fall 2016, EECE 5644, Homework #3

Prof: Jennifer Dy, Deniz Erdogmus **TA: Sadegh Salehi, Hongfu Liu**

Make sure you read the problem statements and answer the questions. Submit answers and plots both online and as a hardcopy. Submit your code to BB. Since this homework requires a lot of plotting, write your code modularly so that making plots with different datasets is easy. For plots on datasets with more than 2 features, plot only the FIRST TWO features (x-y scatter‘ plot)

2.1 (30 pts) Parameter estimation

(a) (5 pts) Estimate the mean and covariances for the three classes in datasetP2Train1.mat. You can assume ML estimation for the Gaussian case, although it is not necessary since the formula for sample mean and covariance are the same independent of the pdf.

(b) (5 pts) Using the shrinkage equation (Eq. 76 from Chapter 3), write a program that uses α to regularize the individual covariance matrices. Plot the training error (percent of misclassification) as a function of $0 < \alpha < 1$. The classification should be done with the gaussian discriminant function from homework 1. Use the function provided. The common covariance should be the covariance estimated from the complete dataset.

(c) (5 pts) Compute the classification error for datasetP2Test.mat for $0 < \alpha < 1$ using the estimated parameters from (a) and (b). What can you say about the difference between plot (b) and plot (c)? What α yields best test performance?

(d) (5 pts) Estimate the mean and covariances for the three classes in datasetP2Train2.mat

(e) (5 pts) Using the shrinkage equation (Eq. 76 from Chapter 3), write a program that uses α to regularize the individual covariance matrices. Plot the training error (percent of misclassification) as a function of $0 < \alpha < 1$. The classification should be done with the gaussian discriminant function from homework 1. Use the function provided.

(f) (5 pts) Compute the classification error for datasetP2Test.mat for $0 < \alpha < 1$ using the estimated parameters from (d) and (e). What can you say about the difference between plot (b) and plot (c)? What α yields best test performance? What is the difference between part (c) and part (f)? Why is shrinkage useful?

2.2 (70 pts) K-means

(a) (20 pts) Implement the basic version of the K-means algorithm as described in the text in Section 10.4.3 on page 527. The implementation should be a function that takes the following inputs:

i. inputData: nSamples x dDimensions array with the data to be clustered

ii. numberOfClusters: number of cluster for algorithm

iii. stopTolerance: convergence criteria

iv. numberOfRuns: number of times the algorithm will run with random initializations

The function should output the results for the run that gave minimum MSE. The outputs should be:

i. estimatedLabels: nSamples x 1 vector with estimated cluster assignment

ii. estimatedMeans: numberOfClusters x dDimensions array with the estimated cluster means

iii. MSE: nIterations x 1 vector with MSE as a function of iteration number

(b) (10 pts) For dataset1.mat, using the true number of clusters (K=2), run the K-means algorithm. Make a 1x3 figure (with subplot) to generate the following plots: (a) true clustering, (b) results of clustering with the estimated means marked as well (c) MSE as a function of iteration number.

(c) (10 pts) For dataset2.mat, using the true number of clusters (K=3), run the K-means algorithm. Make a 1x3 figure (with subplot) to generate the following plots: (a) true clustering, (b) results of clustering with the estimated means marked as well (c) MSE as a function of iteration number. How does K-means perform?

(d) (10 pts) For dataset3.mat, using the true number of clusters (K=2), run the K-means algorithm. Make a 1x3 figure (with subplot) to generate the following plots: (a) true clustering, (b) results of clustering with the estimated means marked as well (c) MSE as a function of iteration number. The scatter plot must be done with dimension 1 on the horizontal axis and dimension 2 on the vertical axis.

(e) (10 pts) For dataset4.mat, using the true number of clusters (K=3), run the K-means algorithm. Make a 1x3 figure (with subplot) to generate the following plots: (a) true clustering, (b) results of clustering with the estimated means marked as well (c) MSE as a function of iteration number.

(f) (10 pts) For dataset5.mat, there is no true number of clusters. Run the K-means algorithm from K=2 to K=15. Get the MSE for each K and plot it. You should see what is called ”the knee.” This would be a K for which the rate at which the within-cluster MSE is decreasing sharply reduces. This suggests we should use the value for K where this knee occurs. Use this K to make a 1x3 figure (with subplot) to generate the following plots: (a) true clustering, (b) results of clustering with the estimated means marked as well (c) MSE as a function of iteration number.

(g) (20 pts) Bonus: implement the K-means++ algorithm with similar inputs to those of part (a). Run this code with the 4 datasets to generate the plots from (b) - (e). Comment on its the convergence when compared to the standard K-means. K-means++ documentation is on BB