Akshay Kenchappa Manjunath

EE 559

Homework 1

January 31, 2020

Problem 1

Introduction

In this problem we implement a nearest-means classifier where we develop a minimum distance to class mean classifier on three given datasets namely synthetic1, synthetic2, and wine dataset. We train the classifier using training dataset as input that are labelled according to class. We develop a code which calculates the class means and utilize them to classify data points. Next, we utilize the provided plotting function to plot the training data points, resulting class means, decision boundaries and decision regions.

After training the classifier, we calculate the classification error rate on both the training dataset and the test dataset. This is expressed as the percentage of the number of misclassified samples to the total number of data samples.

Part (a)

1) Synthetic1 Dataset

Upon training the classifier, we plot the (training-set) data points, the resulting class means, decision boundaries, and decision regions. The figure below depicts the same.

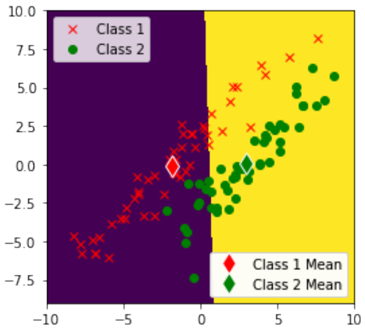

Figure 1: Plot showing classification of Synthetic1 training dataset

1. The purple shaded area is the region which contains training samples of Class 1 and the training data samples belonging to class 1 are represented as red ‘x’.

2. The yellow shaded area is the region which contains training samples of Class 2 and the training data samples belonging to class 2 are represented as green ‘o’

3. The class means of each class is represented by a solid filled diamond.

4. The decision boundary comprises of a line that separates the two decision regions.

The error-rate for Synthetic1 dataset is calculated for the training as well as the test data. The number of misclassified samples for the training data is 21 and the number of misclassified samples for the test data is 24. Consequently, the error-rate for training data is 21% and the error-rate for test data is 24%.

2) Synthetic2 Dataset

Upon training the classifier, we plot the (training-set) data points, the resulting class means, decision boundaries, and decision regions. The figure below depicts the same.

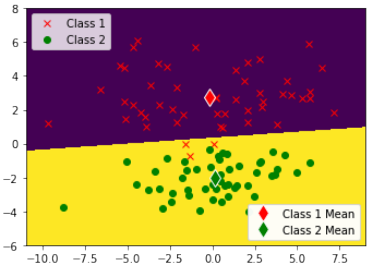

Figure 2: Plot showing classification of Synthetic2 training dataset

1. The purple shaded area is the region which contains training samples of Class 1 and the training data samples belonging to class 1 are represented as red 'x'.
2. The yellow shaded area is the region which contains training samples of Class 2 and the training data samples belonging to class 2 are represented as green 'o'
3. The class means of each class is represented by a solid filled diamond.
4. The decision boundary comprises of a line that separates the two decision regions.

The error-rate for Synthetic2 dataset is calculated for the training as well as the test data. The number of misclassified samples for the training data is 3 and the number of misclassified samples for the test data is 4. Consequently, the error-rate for training data is 3% and the error-rate for test data is 4%.

Part (b)

From the results of the previous section we see that there is quite a difference between the error rates of the synthetic1 and synthetic2 datasets. There are more misclassifications in the case of the synthetic1 dataset. The possible reason for a higher error-rate could be the higher correlation between the features of the synthetic1 data. Hence the designed classifier works better on the synthetic2 dataset.

Part (c)

We now run the classifier for the wine data set. Here we consider the first two features and run the classifier on the training as well as test data points.The figure below shows the plot of the training samples for the wine dataset for the feature values 1 and 2. We have 3 classes which are depicted by the shaded regions and their respective decision boundaries.

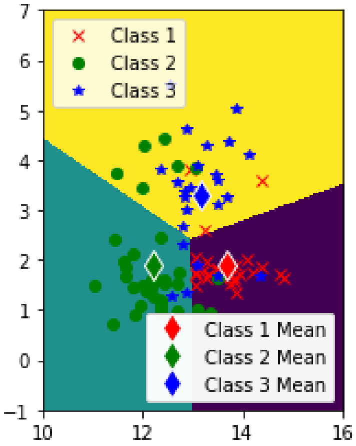

Figure 3: Plot showing classification of Wine training dataset

The error-rate for the wine dataset is calculated for the training as well as the test data. The number of misclassified samples for the training data is 18 and the number of misclassified samples for the test data is 20. Consequently, the error-rate for training data is 20.224719101123593% and the error-rate for test data is 22.47191011235955%.

Part (d)

In this problem we aim to find the best 2 features out of 13 which achieves the minimum classification error. In order to do this I have considered the pairwise combinations of all the 13 features and run the classifier on the training data and determined the error rate on the training data set for all the combinations.

On analysis of the result the feature combination of 1,12 gave the minimum error - rate on the training data.

The figure below shows the plot of the training samples for the wine dataset for the feature values 1 and 12.

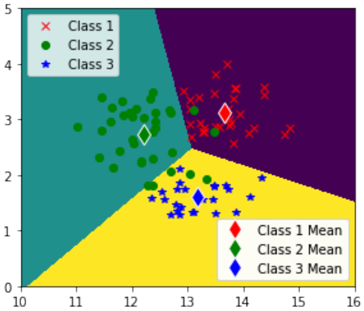

Figure 4: Plot showing classification of Wine training dataset

The error-rate for the wine dataset is calculated for the training as well as the test data. The number of misclassified samples for the training data is 7 and the number of misclassified samples for the test data is 10. Consequently, the error-rate for training data is 7.865169% and the error-rate for test data is 11.235955056179774%.

Part (e)

The combination of different features gives varied range of error rates. The error rates gives us a way of choosing the best features of the 13 for classification. The feature combination which gives the minimum error rate is the best. On running the classifier for a set of different features I got the below results. This table shows that error rate is different for different features.

Feature1	Feature2	Error-rate on training data (%)	Error-rate on test data (%)
1	12	7.8651	12.3595
1	7	8.9887	11.2359
3	7	14.6067	22.4719
1	5	56.1797	44.9438