CS6690- Pattern Recognition

Report on

Programming Assignment 1

Submitted By

Group IV

Divya (ED11S002)
Swarna (EE11S019)
Archana (AM11S027)

9th September 2011

INTRODUCTION

The given training data consists of two classes, each of 400 data points.

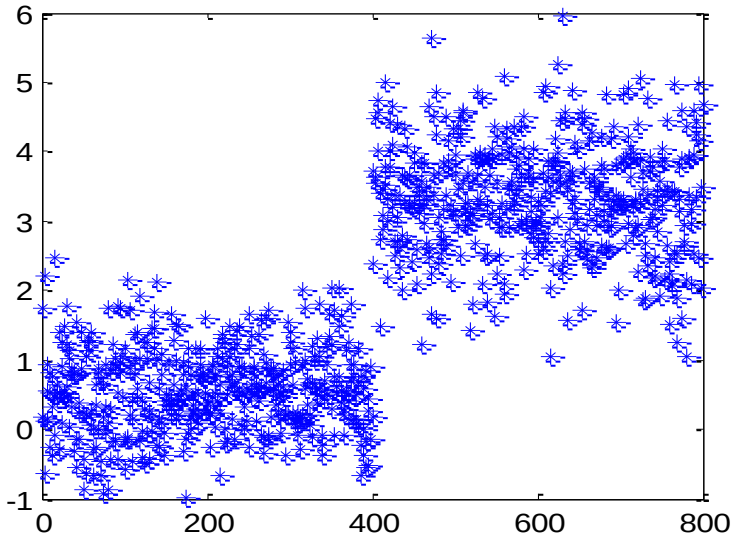


Fig 1. Training Data

The test data also consists of two classes, each of 100 data points.

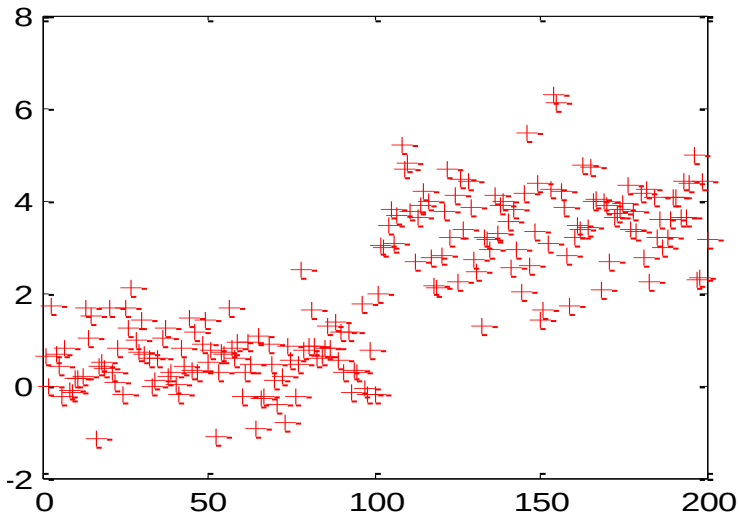


Fig 2. Test Data

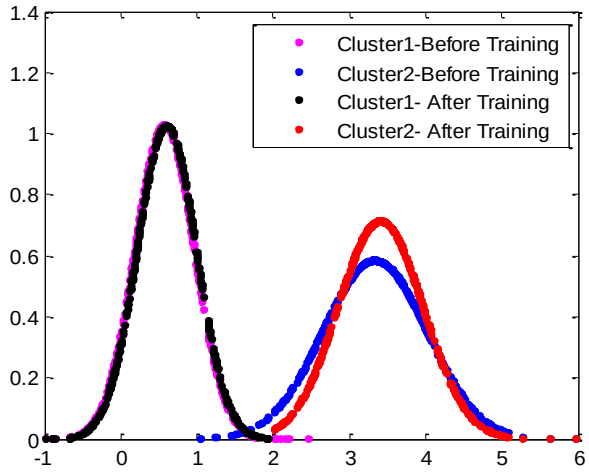
The training data has been classified based on Linde Buzo Gray Algorithm and k-means algorithm using Euclidean distance and Mahalanobis distance.

Classification of the test data has been done by calculating the Mahalanobis distance between the test point and the means of the two clusters from the training data.

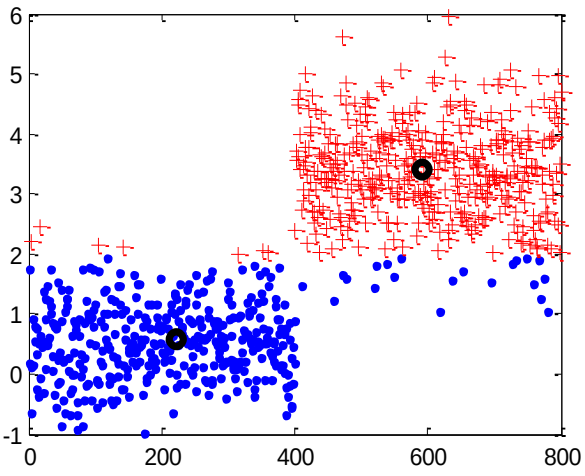
Linde Buzo Gray Algorithm:

a) Using Euclidean norm:

Training data:

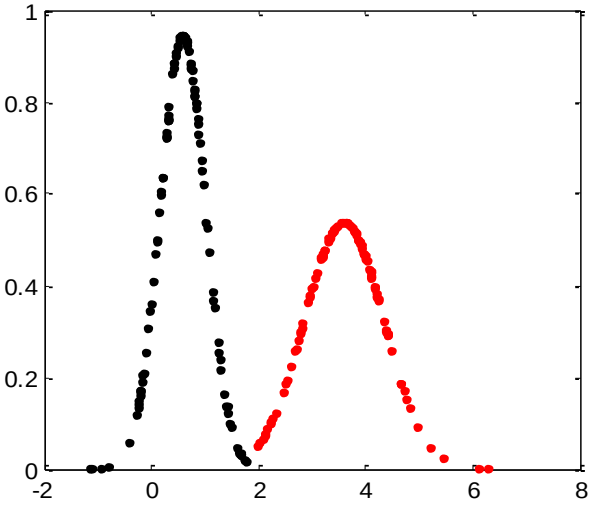


Distributions before & after training

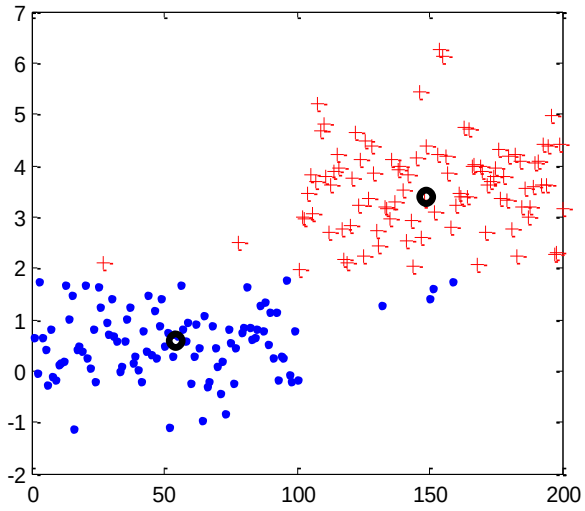


Two clusters and their means

Test data:



Distributions of the two clusters



Two clusters and their means

The confusion matrix for the Euclidean distance:

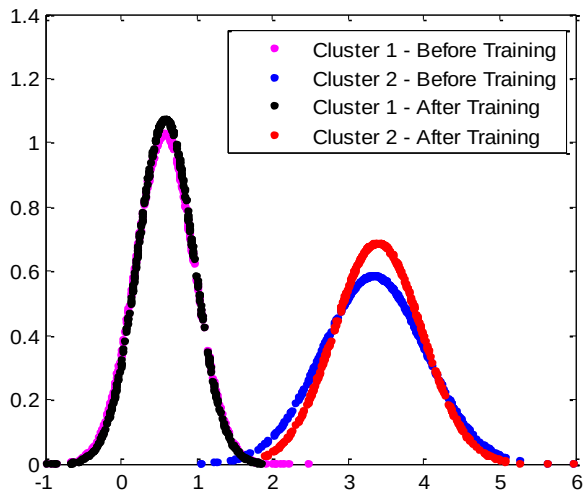
$$W1=\begin{matrix} 98 & 4 \\ 2 & 96 \end{matrix}$$

Accuracy = 97%

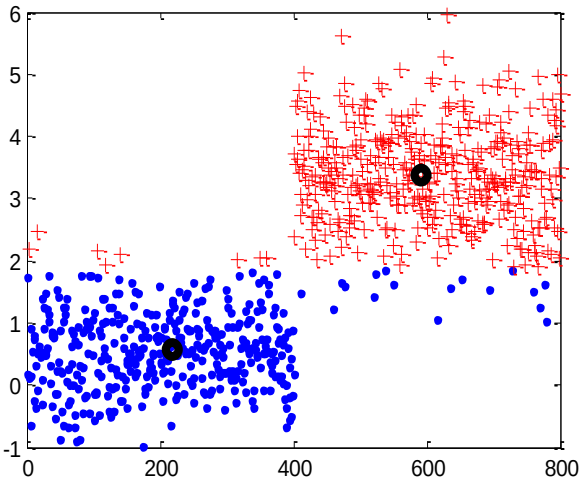
True Positive Rate = 98% False Positive Rate = 4%

b) Using Mahalanobis distance:

Training data:

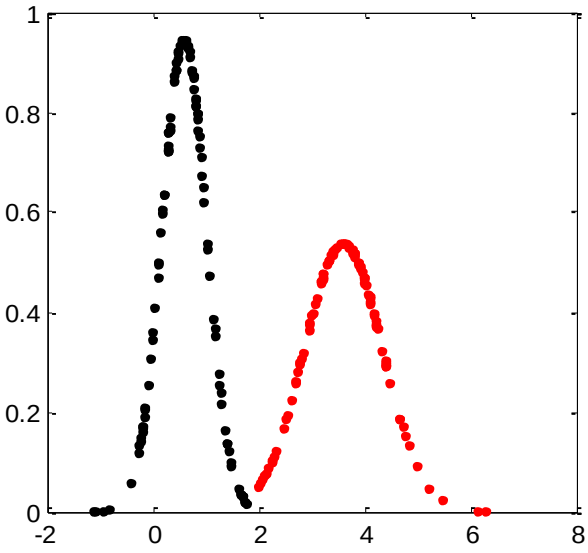


Distributions before & after training

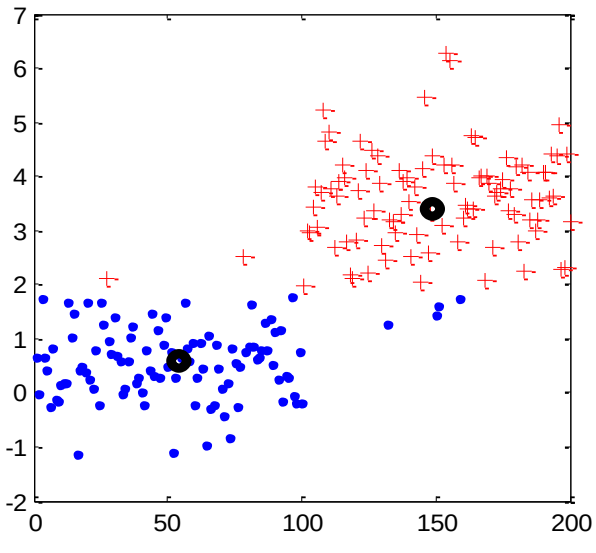


Two clusters and their means

Test data:



Distributions of the two clusters



Two clusters and their means

The confusion matrix for the Mahalanobis distance:

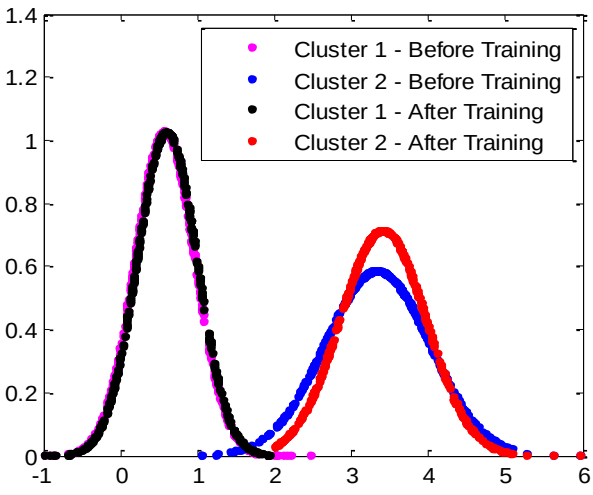
$$W1=\begin{matrix} 98 & 4 \\ 2 & 96 \end{matrix}$$

Accuracy=97%

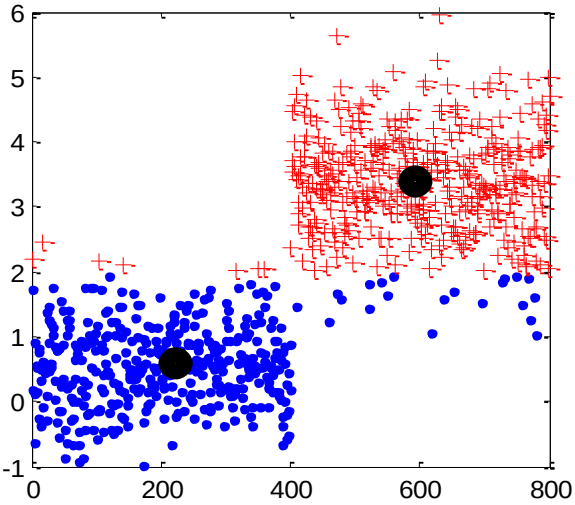
True Positive Rate = 98% False Positive Rate = 4%

3) K - Means Algorithm

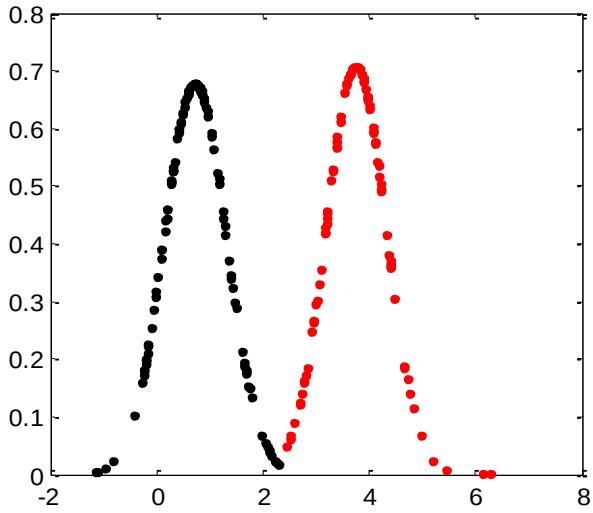
a) Using Euclidean norm:



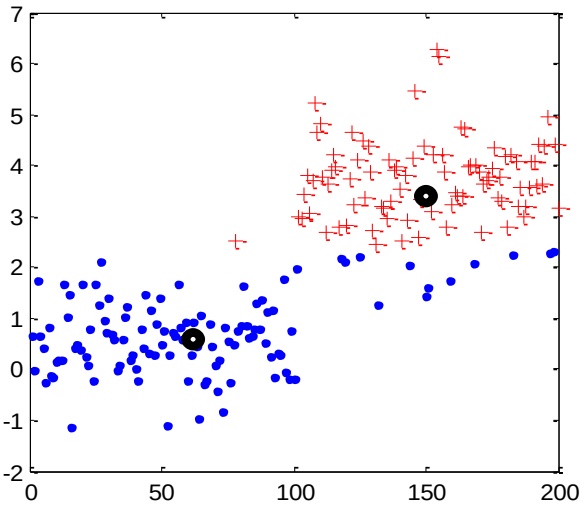
Distributions before & after training



Two clusters and their means



Distributions of the two clusters



Two clusters and their means

The confusion matrix for the data using k-means with training data classified based on Euclidean distance:

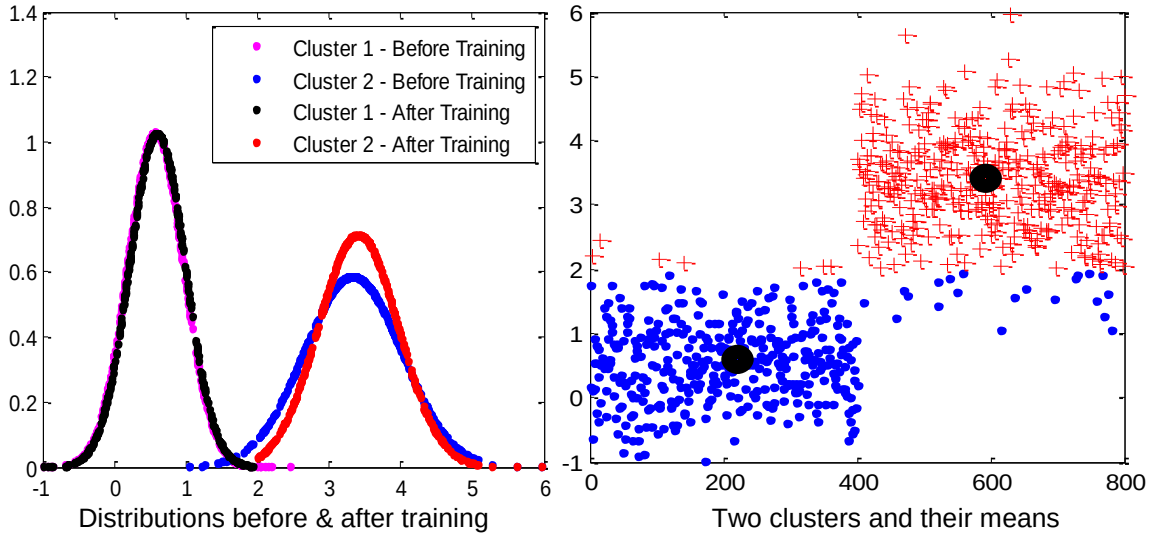
$$W3 = \begin{matrix} 99 & 13 \\ 1 & 87 \end{matrix}$$

Accuracy = 93%

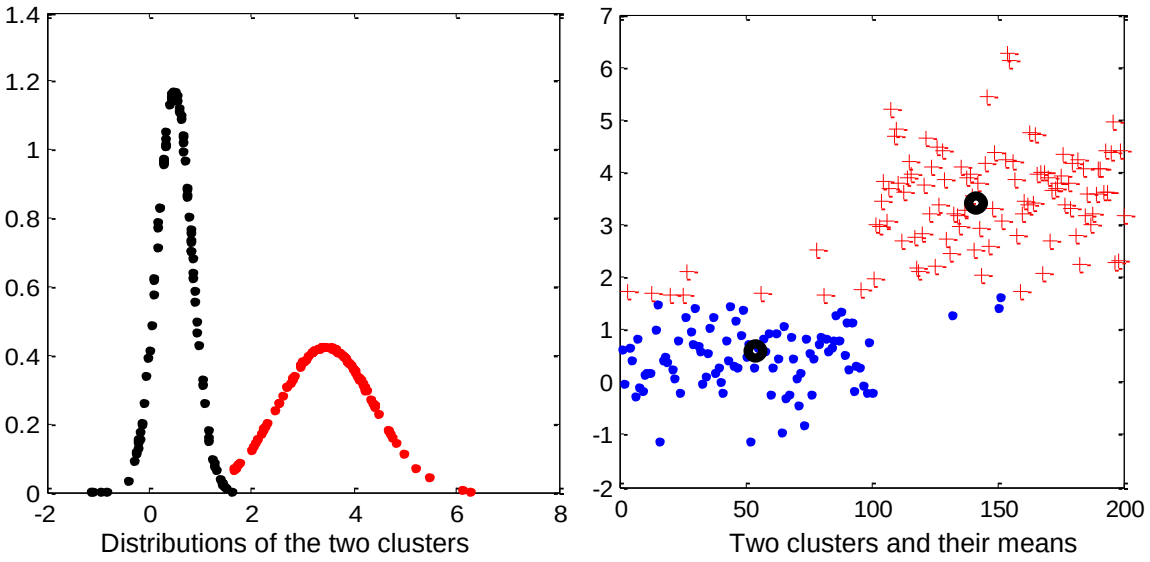
True Positive Rate = 99% False Positive Rate = 4%

b) Using Mahalanobis distance:

Training data:



Test data:



The confusion matrix for the data choosing k-means with training data classified based on mahalanobis distance:

$$W4=\begin{matrix}91 & 3\\ 9 & 97\end{matrix}$$

Accuracy = 94%

True Positive Rate = 91%

False Positive Rate = 3%

OBSERVATIONS

The data sets were classified using Linde Buzo Gray Algorithm and k-means algorithm by calculating Euclidean distance and Mahalanobis distance. The accuracies, True Positive Rate and False Positive Rate were as follows:

		Accuracy	TPR	FPR
Linde Buzo Gray Algorithm	Euclidean distance	97	98	4
	Mahalanobis distance	97	98	4
k-means algorithm	Euclidean distance	93	99	4
	Mahalanobis distance	94	91	3