Pattern Recognition System

Constructing a Bayes Classifier to classify new points

Vinay Mohan Behara

CEG7570- Pattern Recognition

Wright State University

1. Abstract

Pattern Recognition is a study of how machines can observe the environment, learn to distinguish the patterns of interest and make reasonable decisions about the categories of patterns. It categorizes the input data into identifiable classes. Applications of Pattern Recognition Systems are numerous and cover a broad scope of activities. The examples include Speech Recognition in engineering, Stock exchange forecast in economy, classification of rocks in geology and many other.

2. Introduction

The project is divided into three parts. In the first part we remove the outliers and each feature was normalized to have a mean of 0 and a standard deviation of 1. In the second part by PCA (Principal Component Analysis) the number of features are reduced from d to 2 and the Squared and Percent Error are calculated taking the new generated two feature set. In the third part the Bayes Classifier is built and the new points are given as input for the classifier to classify. The Recognition Rate is calculated.

3. Approach

Project Part 1

In order to start the we first import the required data from the main dataset. Here we are using the BankNoteData and Iris datasets. Once the dataset is imported, we should start the process of Normalization. We will find the mean and variance. These values are normalized to 0 and 1 respectively. The normalization removes the outliers.

The formulas used for Normalization:

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots, l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

Now the process of removing outliers takes place. So the process involved is if the normalized observations are greater than 3.0 or smaller than -3.0, we remove those observations in entire feature set. These are considered as outliers and removed. So the entire row in the matrix is removed so that the values are out of outliers. These resulted values form a matrix with the class added to it.

In the next step, the graph is plotted in a two dimensional feature space. This graph is based on the values selected by the user.

Project Part 2

In this part, we will import the normalized values obtained from project part 2. In the process of PCA (Principal Component Analysis), the Covariance Matrix is calculated. PCA removes the linear dependency between features.

The Covariance Matrix is used to calculate the Eigen values and Eigen vectors. The function eigs is used to calculate these values. The two maximum values are taken into Consideration. So, these values are sorted.

Corresponding to the Eigen values, the Eigen vectors are taken. These values are multiplied with the imported data matrix. So, this is the process of reducing the features from d to 2.

After the above process, two new features are obtained. These features are plotted using a graph in two dimension.

In the next phase, the Squared Error and Percent Error are calculated using the following formulas.

Squared Error = Sum of two smallest Eigen Values

Percent Error = ((sum of two smallest Eigen Values)/(sum of all Eigen Values))*100

In the squared Error calculation, the smallest values are obtained when we performed the sort operation to get the larger values in the previous steps. The other values are taken as smaller values. The percent Error should be less than 50 percent for a decent result.

Project part 3

In the third part of the project, we build a Bayes classifier to classify the new points. So, In order to create a minimum distance classifier, we first calculate the mean and standard deviation of the two features obtained in the project part 2 that is from the PCA.

We then normalize the features to have a standard deviation of 1. In this case we will not change its mean of the features.

In the next phase, we will find the mean of points in each class. Now we can take a point and classify it. The input is given by the user and a point (p) which has four features is taken from the user.

The point(p) is classified. In the next phase we normalize this point using the mean and standard deviation from project part 1. The obtained point is taken as q.

In order to get the point r, we use the PCA. We transform the point q using the Principal Component Analysis and the result obtained is taken as the point r. So the point q is multiplied with the eigen vectors of corresponding largest eigen values in project part 2.

In the next phase, we normalize the point r using the standard deviations and this is normalized based on the mean of points obtained after PCA. So in order to do this, we find the two distances between r to mean of two class 1 and r to mean of class 2.

Then we classify the point into either the first class or the second class by using the minimum distance. So the point will be assigned to the first class when the distance from the point to the class is less and the same with class two.

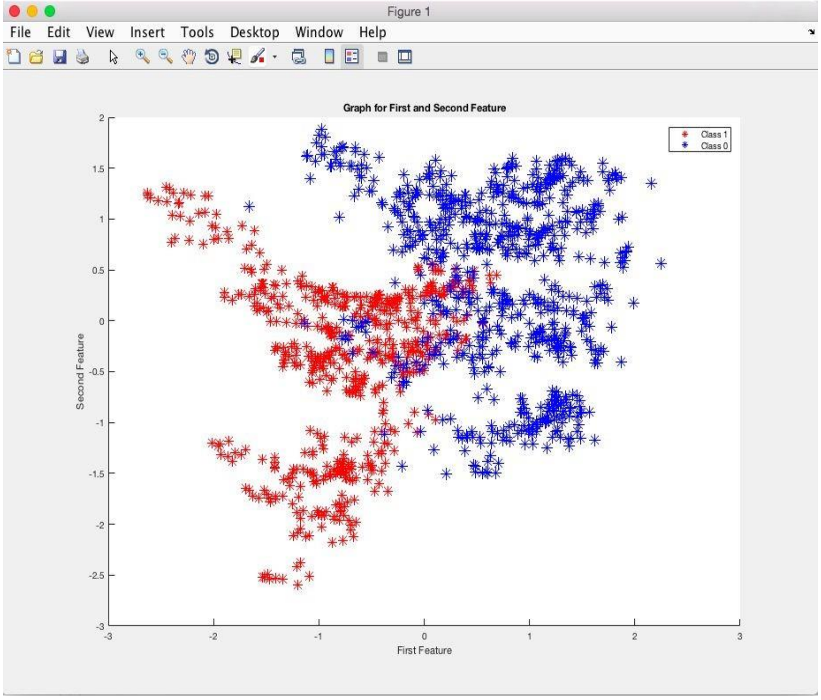
In the next phase we plot a graph with the new point and the two features.

Now we calculate the recognition rate for the total pattern recognition system.

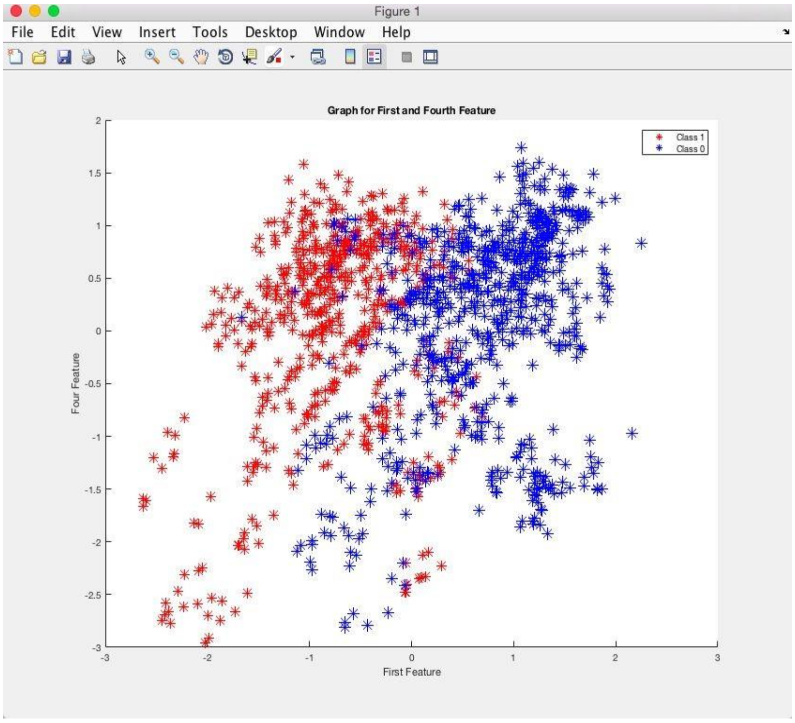
4. Test Cases

Project part 1(BankNoteData)

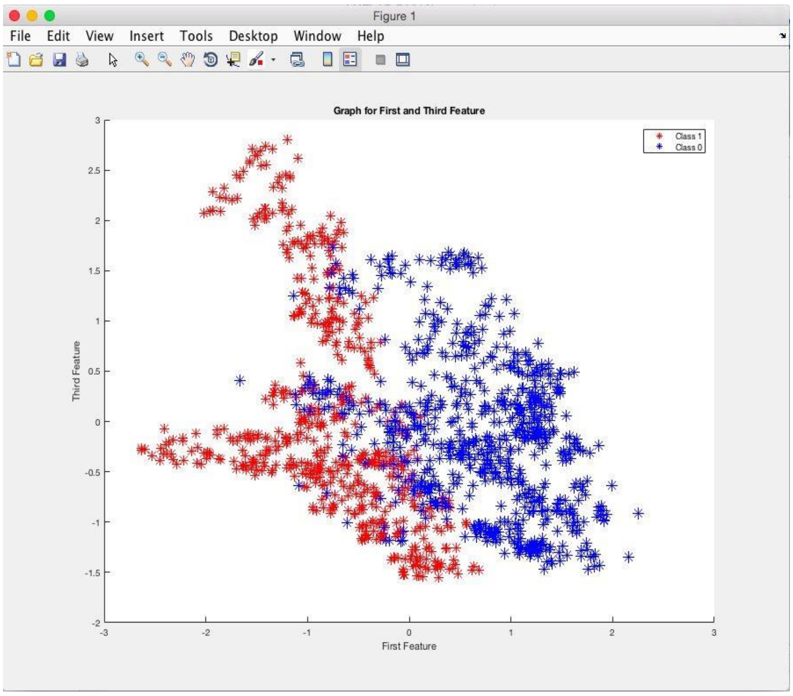
- Output Graph for first feature and second feature



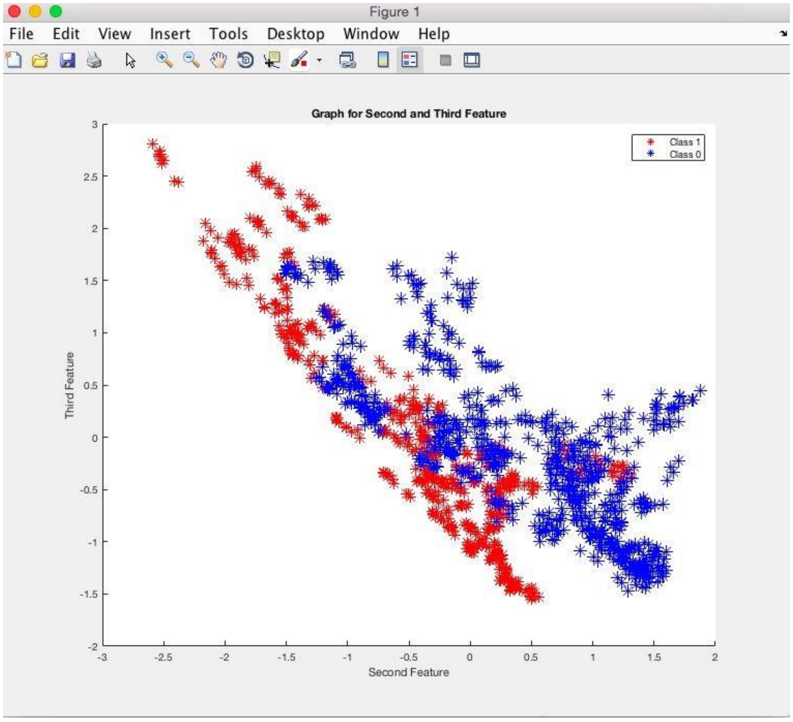
- Output graph for first feature and the fourth feature



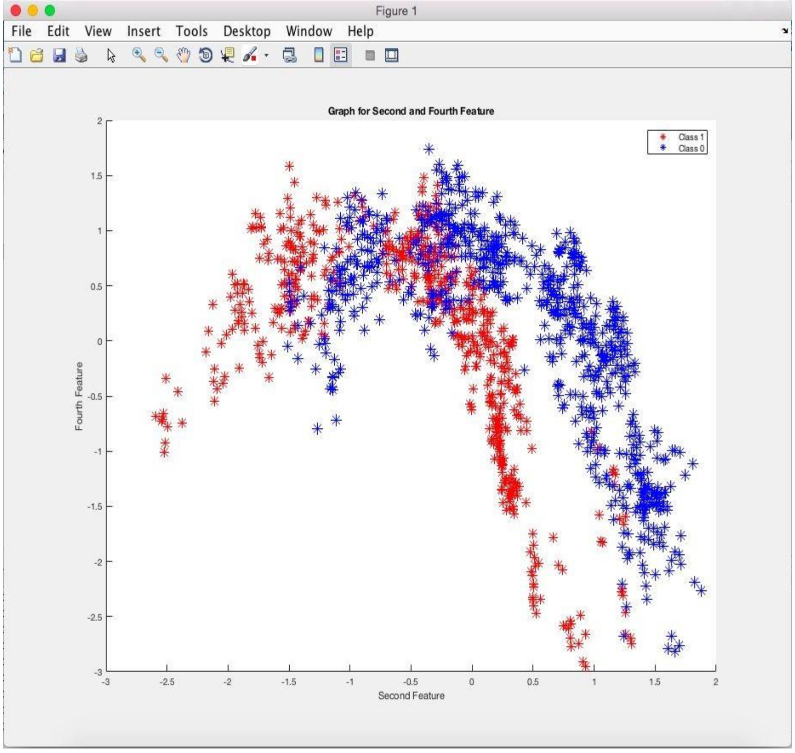
- Output graph for first feature and third feature



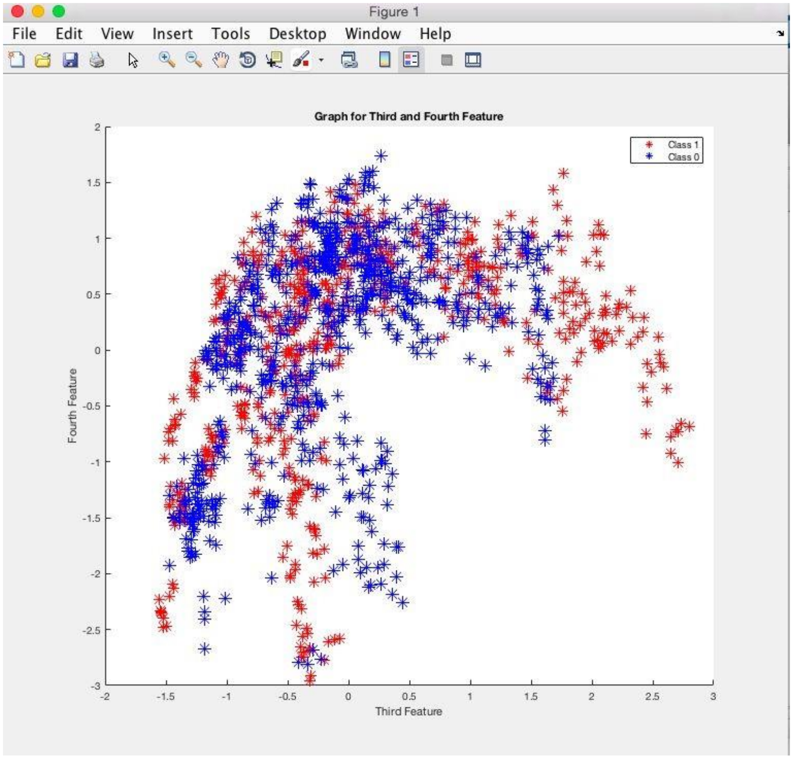
- Output graph for feature second feature and third feature



- Output graph for second feature and fourth feature

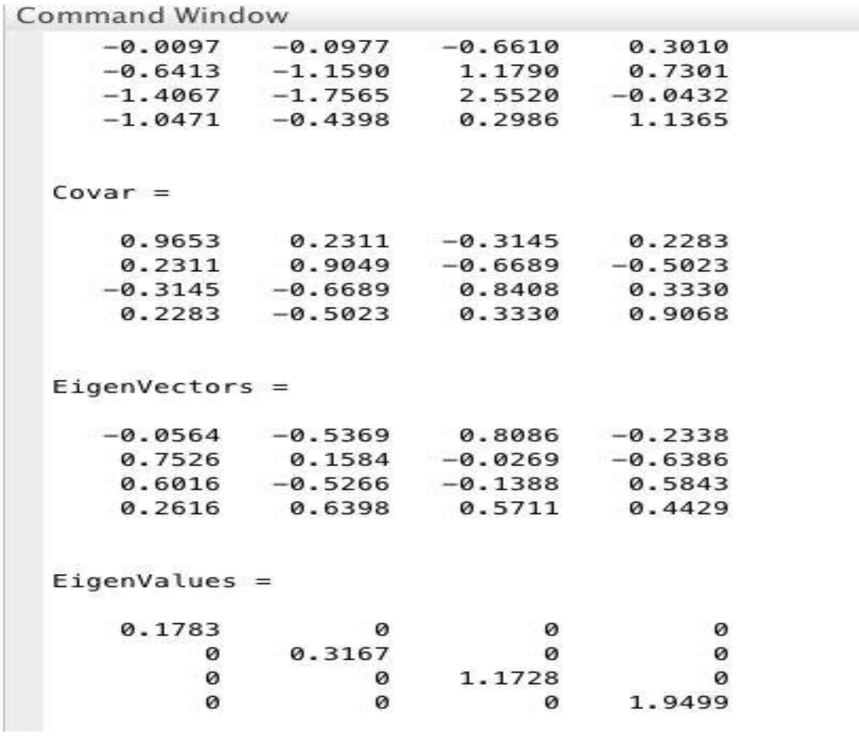


- Output graph for third feature and fourth feature



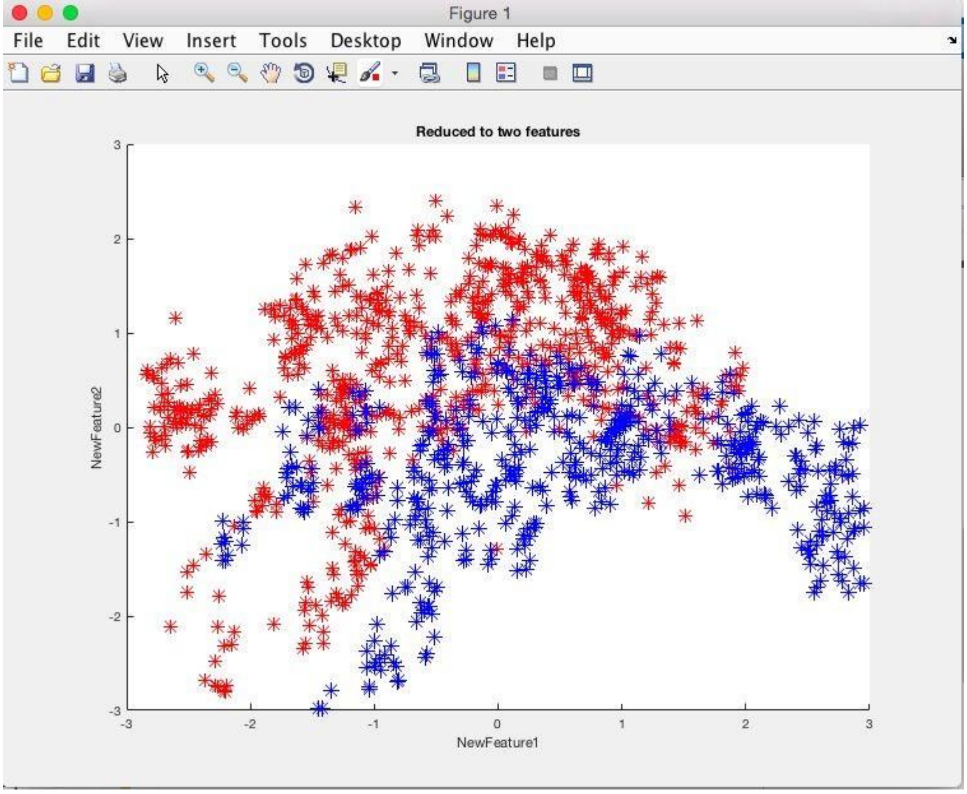
Project part 2

The output of features reduced to two



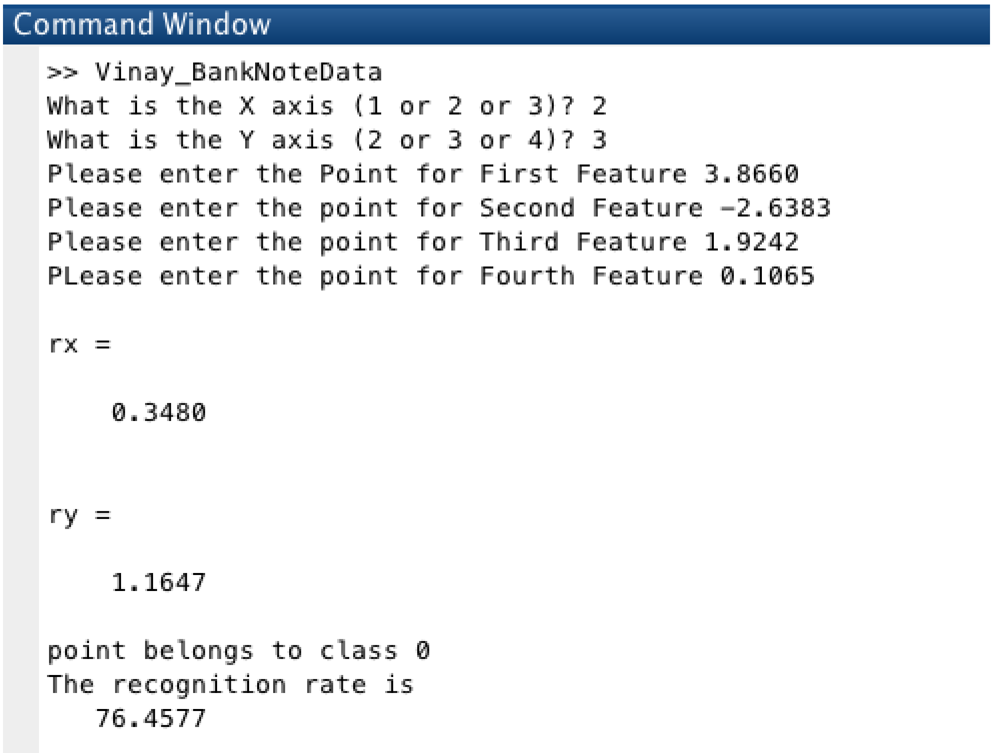
```
   -0.0097   -0.0977   -0.6610    0.3010
   -0.6413   -1.1590    1.1790    0.7301
   -1.4067   -1.7565    2.5520   -0.0432
   -1.0471   -0.4398    0.2986    1.1365


Covar =

    0.9653    0.2311   -0.3145    0.2283
    0.2311    0.9049   -0.6689   -0.5023
   -0.3145   -0.6689    0.8408    0.3330
    0.2283   -0.5023    0.3330    0.9068


EigenVectors =

   -0.0564   -0.5369    0.8086   -0.2338
    0.7526    0.1584   -0.0269   -0.6386
    0.6016   -0.5266   -0.1388    0.5843
    0.2616    0.6398    0.5711    0.4429


EigenValues =

    0.1783         0         0         0
         0    0.3167         0         0
         0         0    1.1728         0
         0         0         0    1.9499
```

The graph for the two features



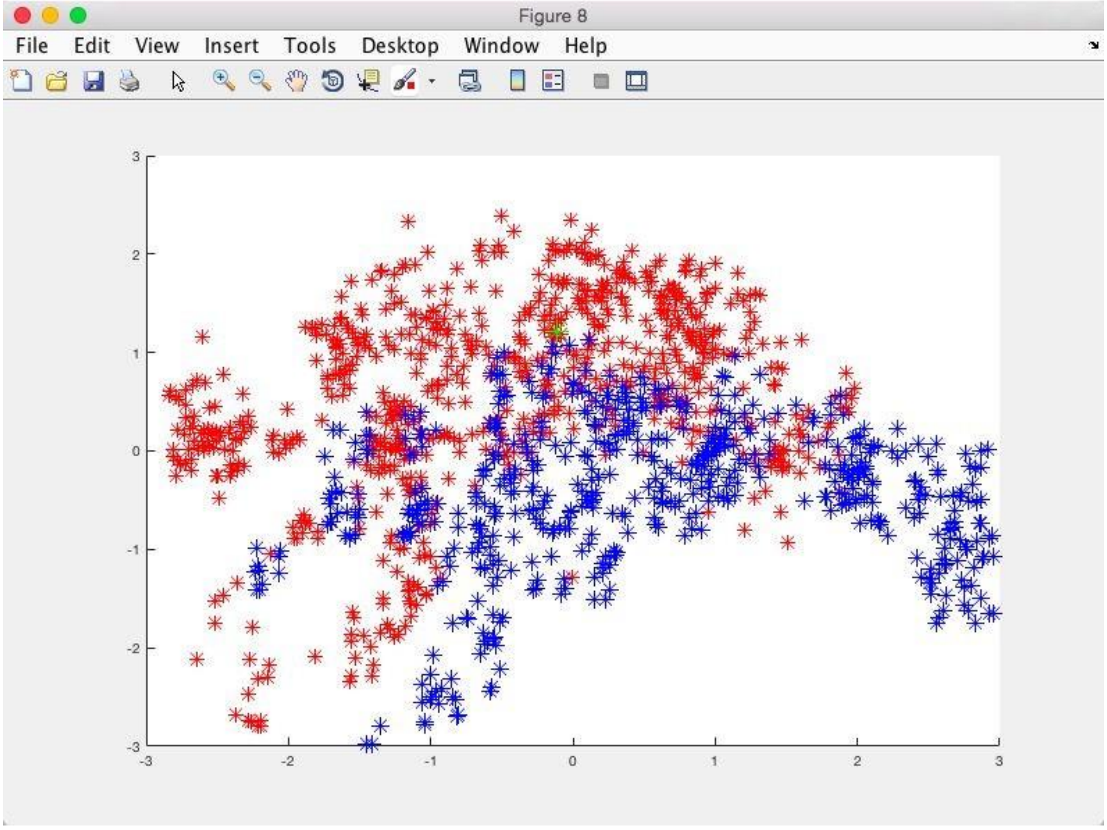
Project part 3

Classification of the point given by the user



```
>> Vinay_BankNoteData
What is the X axis (1 or 2 or 3)? 2
What is the Y axis (2 or 3 or 4)? 3
Please enter the Point for First Feature 3.8660
Please enter the point for Second Feature -2.6383
Please enter the point for Third Feature 1.9242
PLease enter the point for Fourth Feature 0.1065

rx =

    0.3480


ry =

    1.1647

point belongs to class 0
The recognition rate is
   76.4577
```

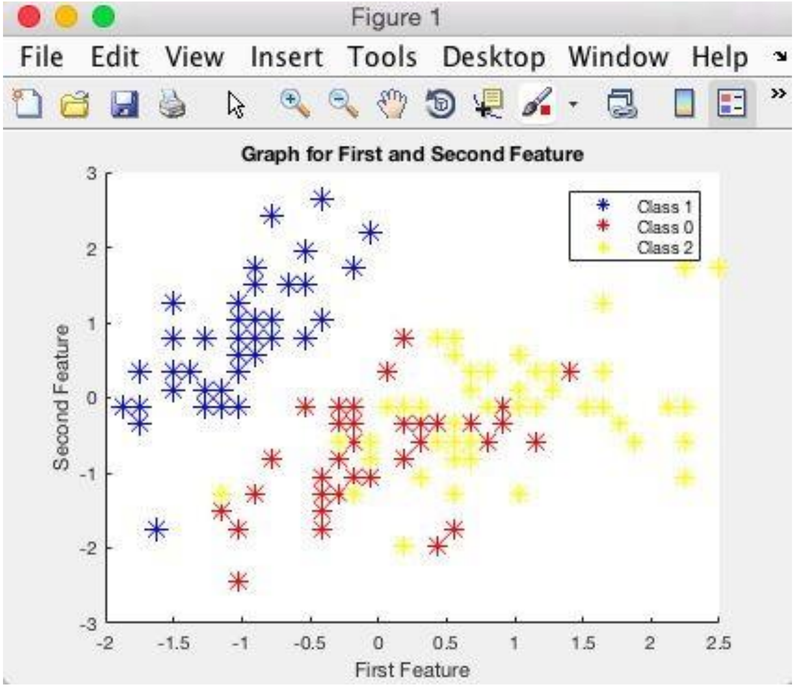
Output graph for the point and the new features



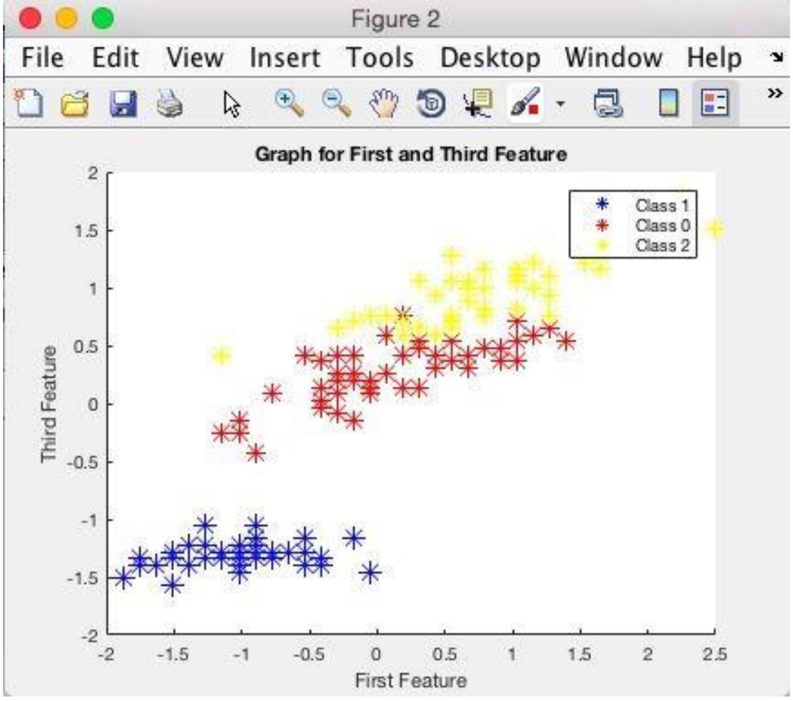
IRIS Dataset

Project Part 1

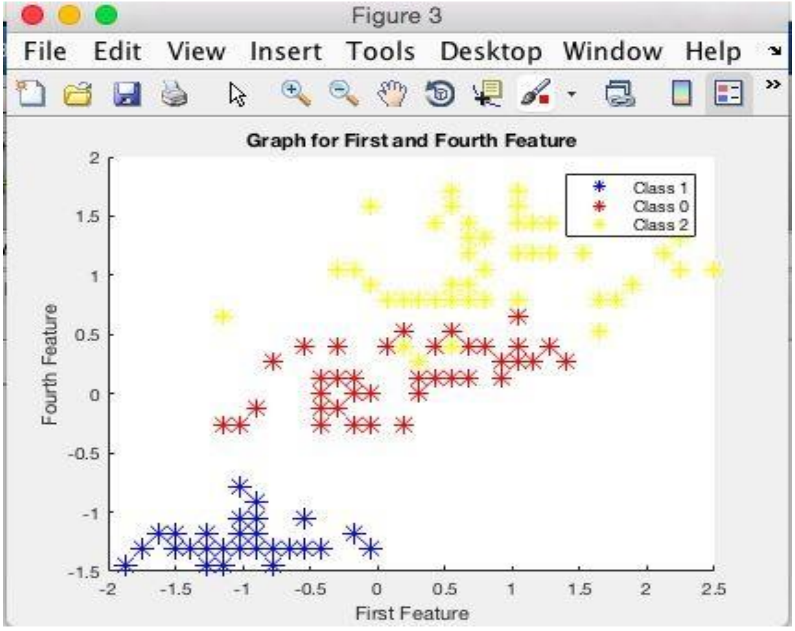
Output Graph for first feature and second feature



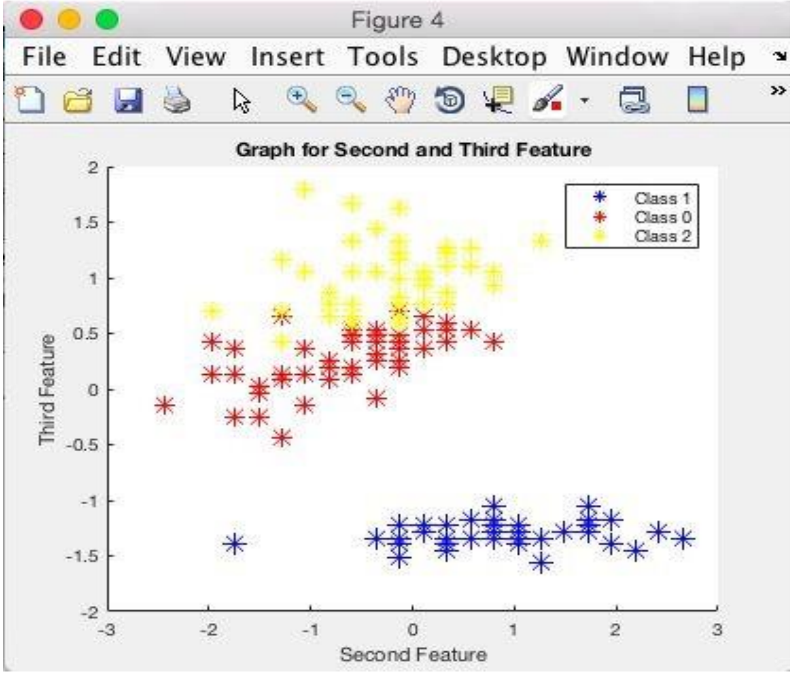
Output graph for first feature and third feature



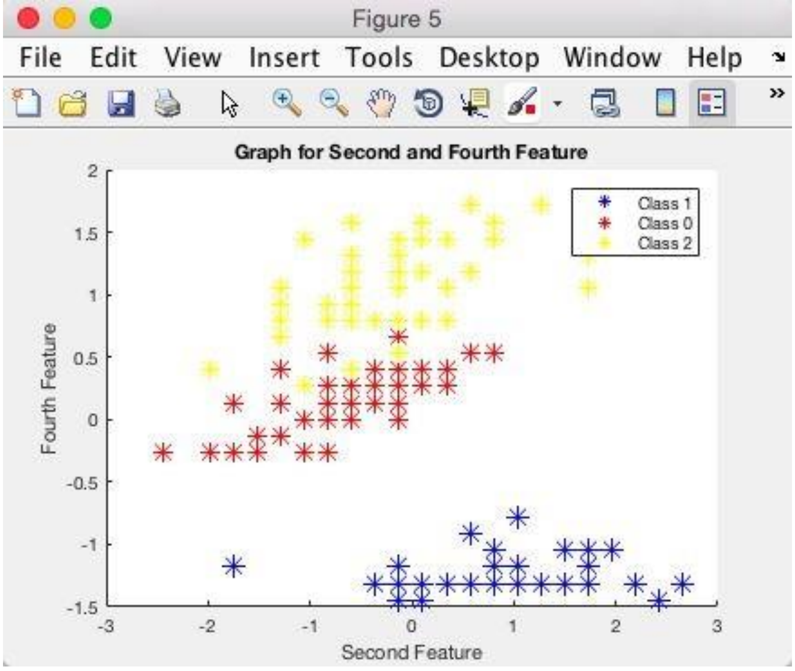
Output graph for feature one and feature four



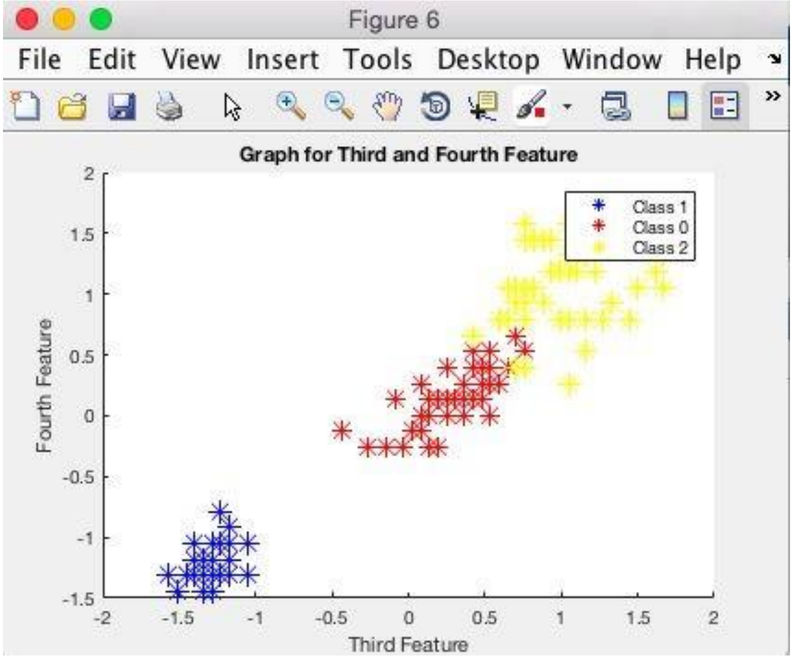
Output graph for feature two and feature three



Output graph for feature two and feature four

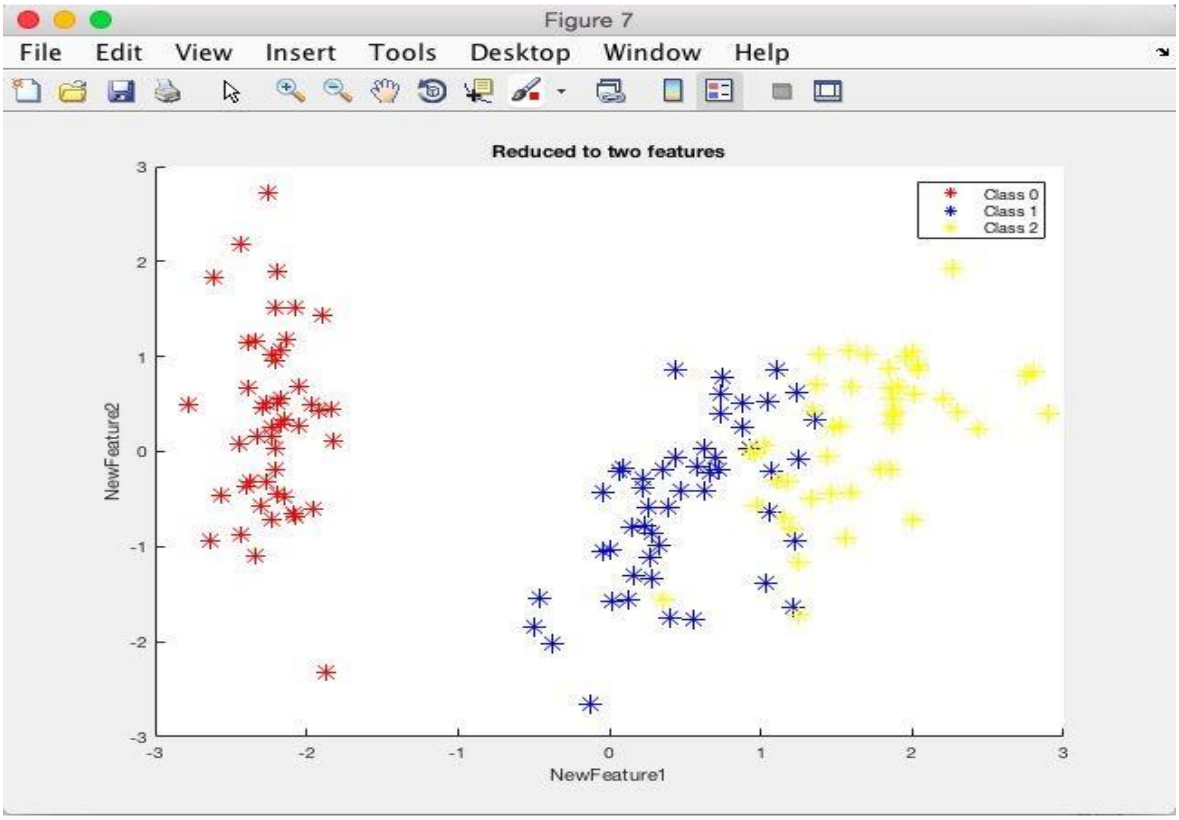


Output graph for feature three and feature four



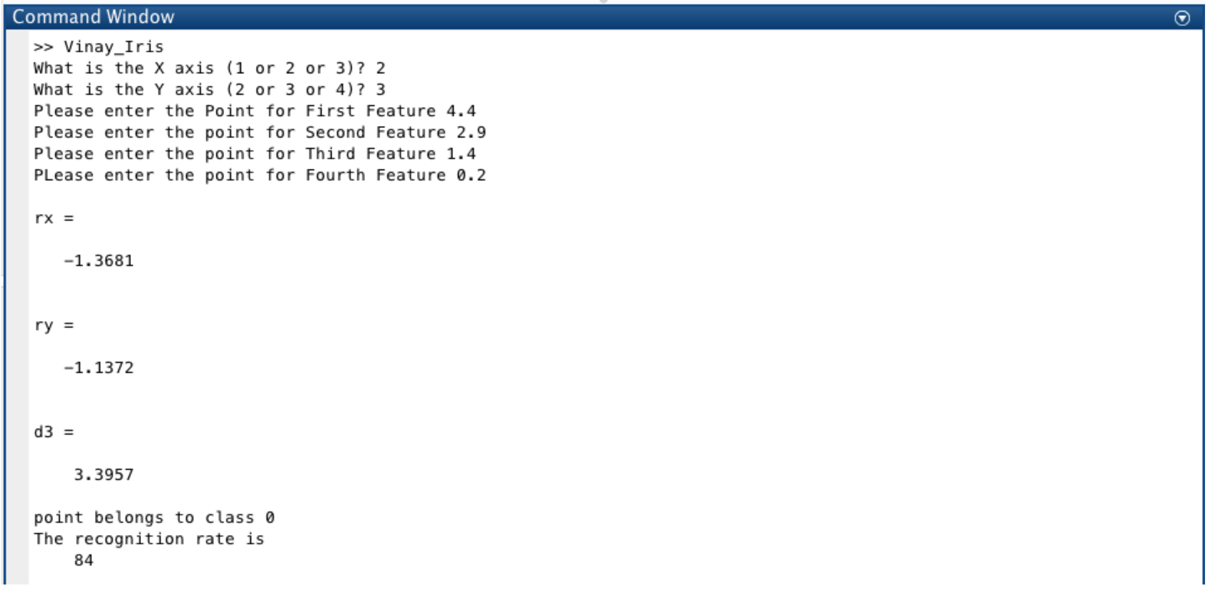
Project Part two

The output of features reduced to two and its graph



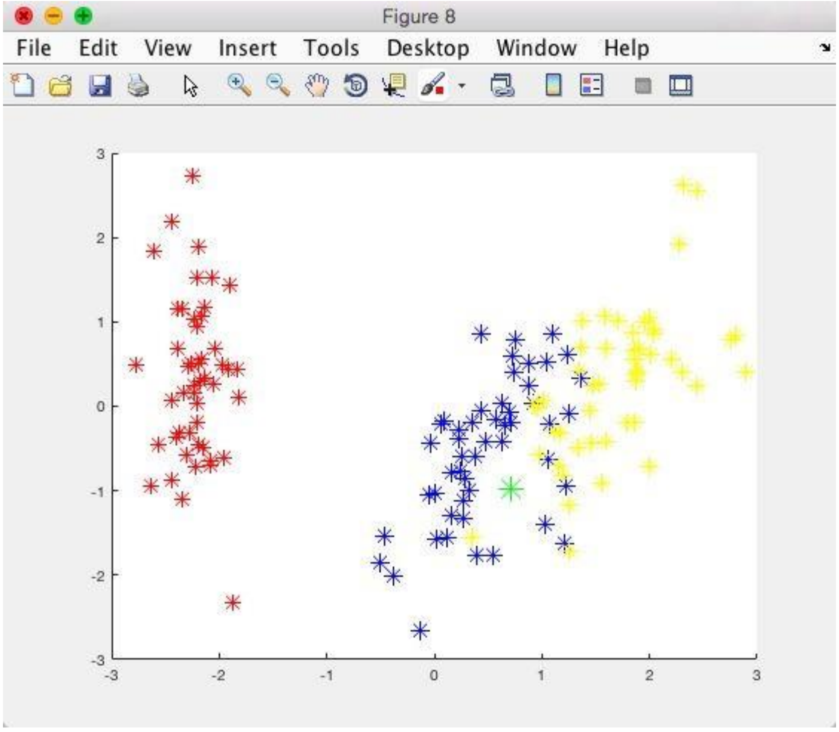
Project Part 3

Classification of point given by the user



```
>> Vinay_Iris
What is the X axis (1 or 2 or 3)? 2
What is the Y axis (2 or 3 or 4)? 3
Please enter the Point for First Feature 4.4
Please enter the point for Second Feature 2.9
Please enter the point for Third Feature 1.4
PLease enter the point for Fourth Feature 0.2

rx =

   -1.3681


ry =

   -1.1372


d3 =

    3.3957

point belongs to class 0
The recognition rate is
    84
```

Output Graph for point and new features



5. Conclusion

The final output of the pattern recognition system gives us the ability to determine and classify the point given by the user into the a particular class based on the process prescribed above. We can determine the class of each point given by the user.

The recognition rate is calculated for the both data sets used. The recognition rate for the first data set BankNoteData is 73%. The recognition rate for Dataset Iris is 84%. This shows the accuracy of the above Pattern recognition System developed for the given two datasets.