Machine Learning Assignment 1
Linear Discriminant Analysis 1

Group Members

Ritesh Kumar Singh	2019H1030154H
Simran Batra	2019H1030024H
Parth Bhope	2019H1030023H

Dataset description and Train test split

The dataset a1_d1.csv contains 1000 labeled data points where a datapoint has two attributes x and y corresponding to the x and y coordinates on a 2D plane respectively The label values are either 0 or 1 implying that the points belong to either class 0 or class 1.
After separating the class 0 and class 1 data points we found that there are 500 data points that belong to class 0 and 500 data points belonging to class 1.

We separated the class 0 and class 1 data into training and testing sets in a 80:20 ratio, And appended the individual training sets to get the complete training dataset.Similarly we appended the test sets. So the training set contains 800 data points 400 points of each class and the testing set contains 200 data points 100 of each class selected randomly.

Dimension reduction

The vector on which all the points will be projected to simplify the classification process is found by the following process.
We used Numpy to find the mean vector for each class the mean vector has dimensions(1x2)
The within class covariance matrices are also found using the function ‘cov()’ in Numpy The within class has dimensions (2x2) as there are 2 input features.
Matrix addition of the two within class covariance matrices gives the total covariance matrix.The inverse of this matrix is now calculated using Numpy’s np.linalg.inv(sw) function.
The multiplication of the inverse matrix (2x2) and the vector obtained after subtracting the mean vectors(2x1) gives us a (2x1) vector which according to fisher’s criterion is the

best vector on which if the points are collapsed will give the best separation in the two classes.

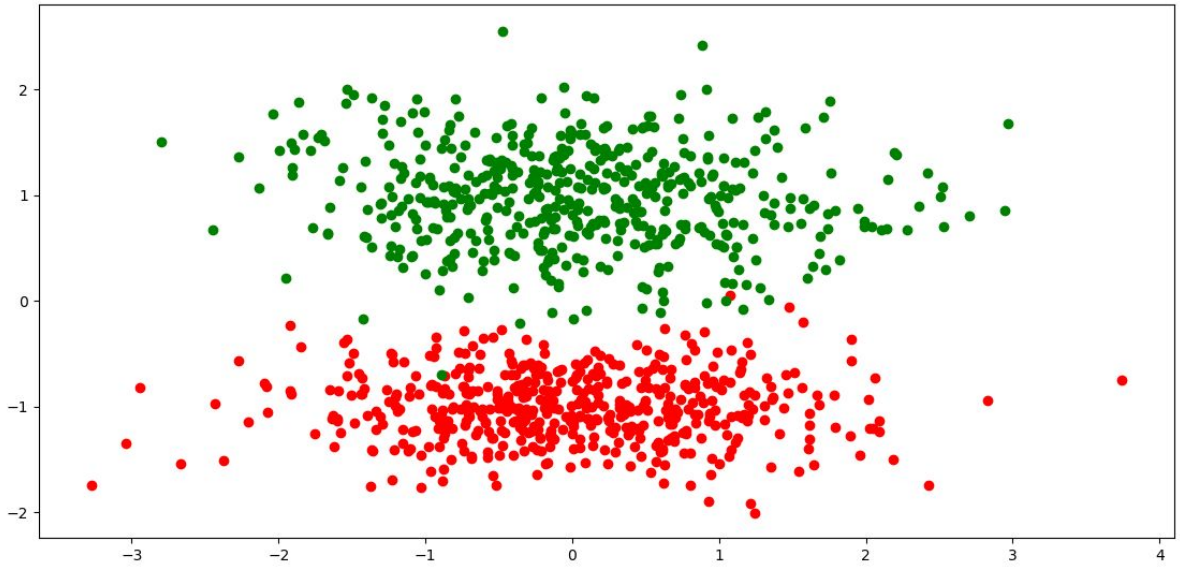
Transforming the features

As we have the best possible vector now, all points are collapsed on this vector and now each data point has only one attribute which is the distance of the point from origin along the vector. The dot product of the data point and the vector gives the transformed value for the data point.

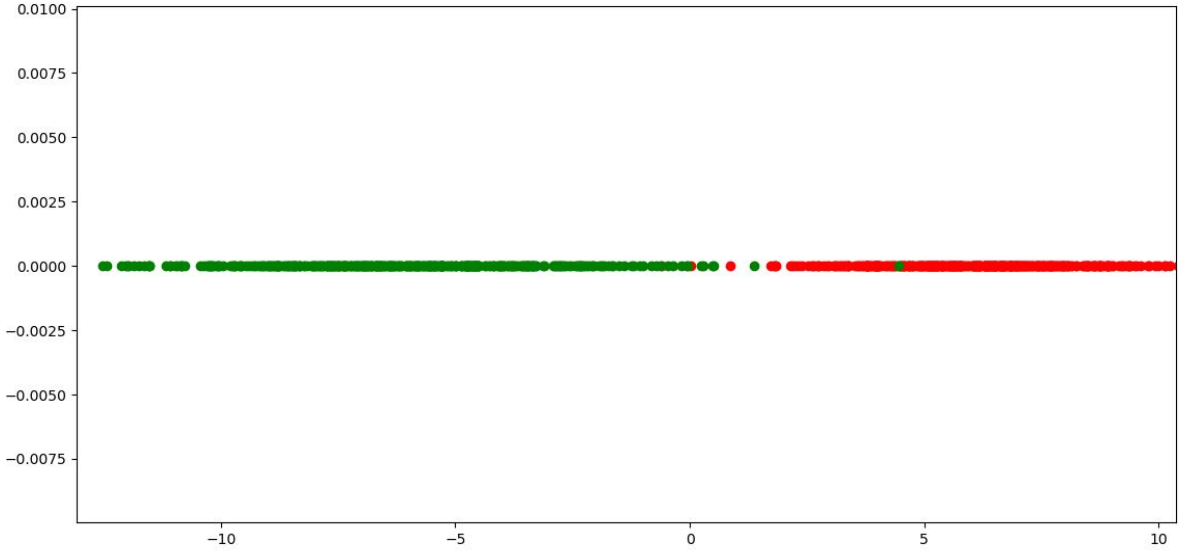
Plotting the points

We used Matplotlib to plot the points

The points in original dimension class 0 are green and class 1 are red points

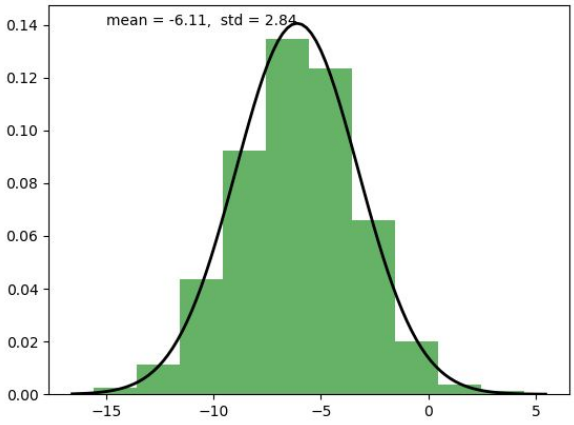


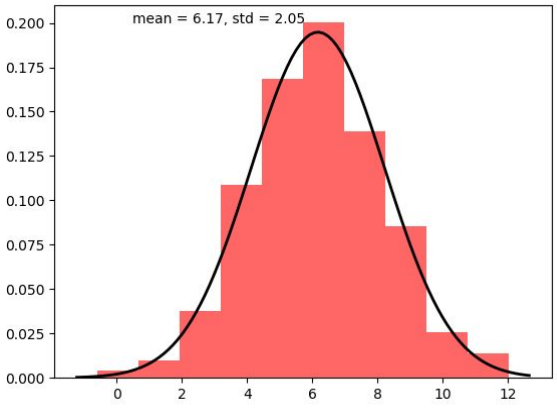
The points after collapsing



Fitting normal distribution

The points are assumed to follow normal distribution so using functions in scipy.norm,normal distribution is fit to the data, the mean and the standard deviation was found using numpy.





The points in class 0 have mean of -6.11 and a standard deviation of 2.84 whereas the points in class 1 have mean 6.17 and a standard deviation of 2.05. The means are well separated.

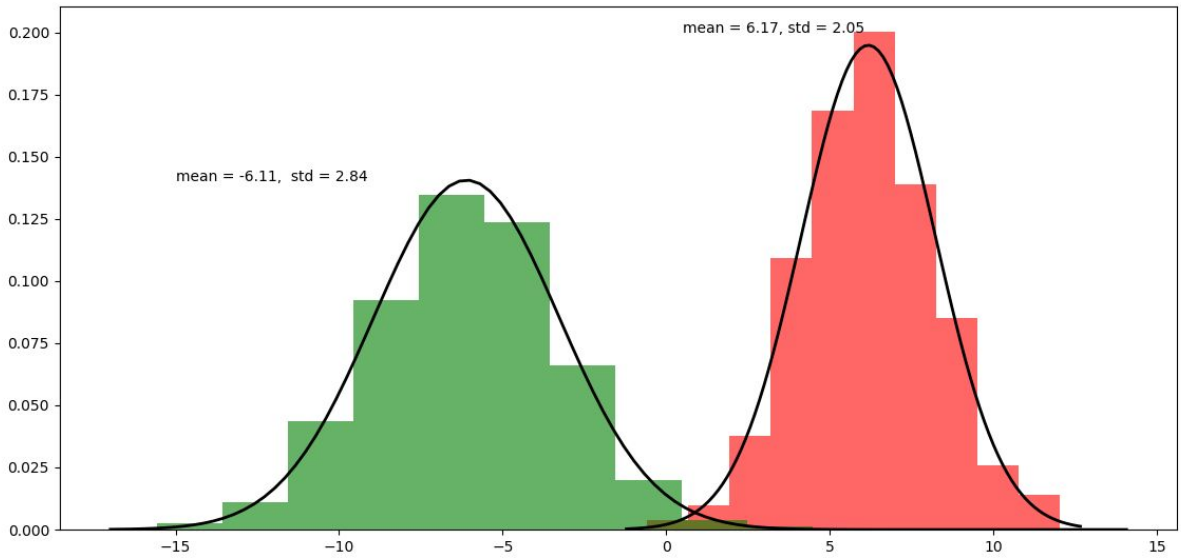
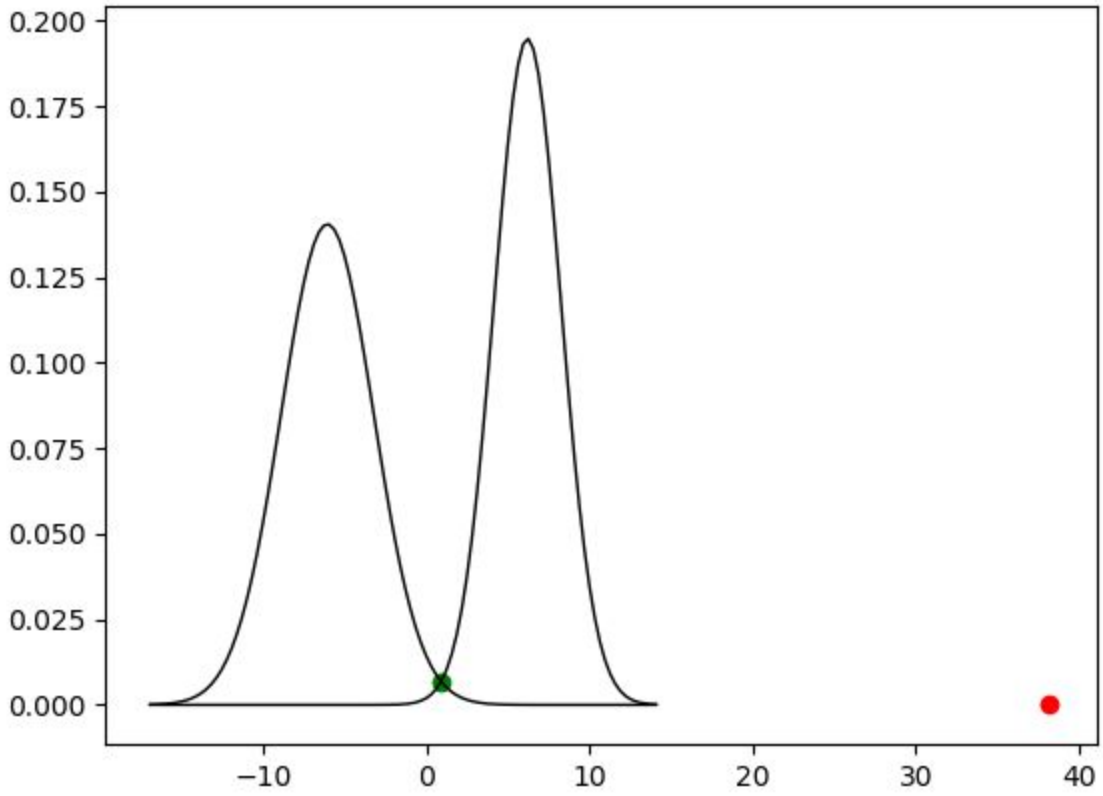


Fig showing the intersection of the two normal curves.



From the fig it is clear that the intersection point is closer to 0 after calculating the intersection points of the two curves which are actually the roots of the quadratic equation obtained by equating the normal curve equations. The roots were obtained as follows x= [38.15229031 0.87141266] ,hence from the fig we concluded as x = 0.87141266 as the discriminating point.

Testing and Results

Every data point from the testing set in the original dimensions, is transformed into a single dimension, which is done by taking the dot product of the data point and the vector.
So if the value obtained after the dot product is less than or equal to 0.8714.. The point is predicted as belonging to class 0 else it is labelled as belonging to class 1.

The testing set has 200 examples with 100 class 0 and 100 class 1 points.

Confusion matrix:

df_confusion - DataFrame

Index	0	1
0	100	0
1	2	98

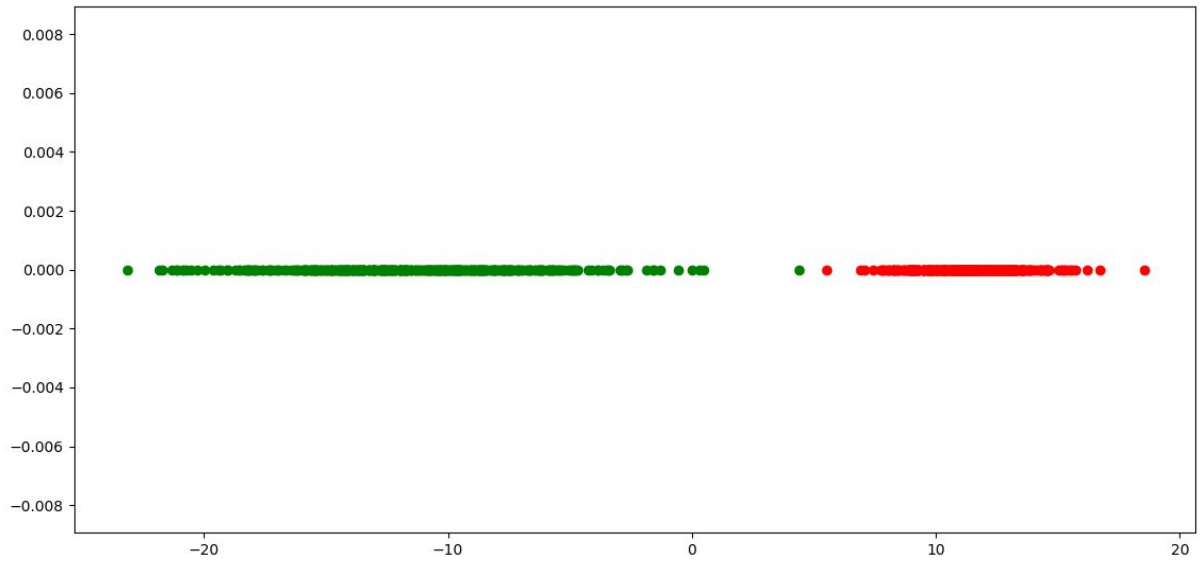
The accuracy obtained was 0.99 and the F score 0.99009.

Machine Learning Assignment 1
Linear Discriminant Analysis 2

Dataset

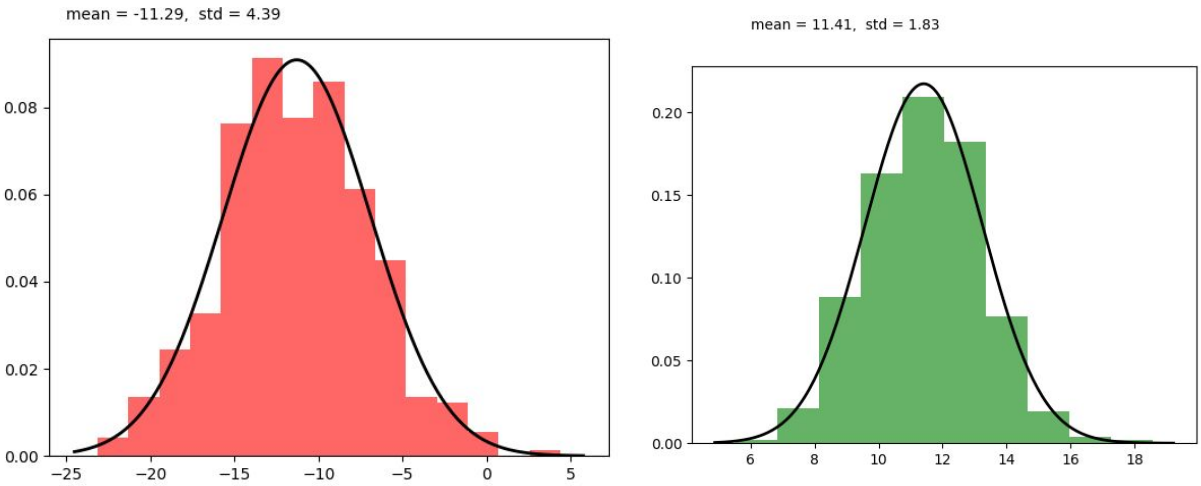
The difference in the part 2 of LDA is that there are 3 input features , so we go from 3D to just 1D for classification. The classes are again the same as before , i.e it is a binary classification problem

The dimensions of the mean vector collapsing vector are (3x1), the covariance matrix has dimensions 3x3



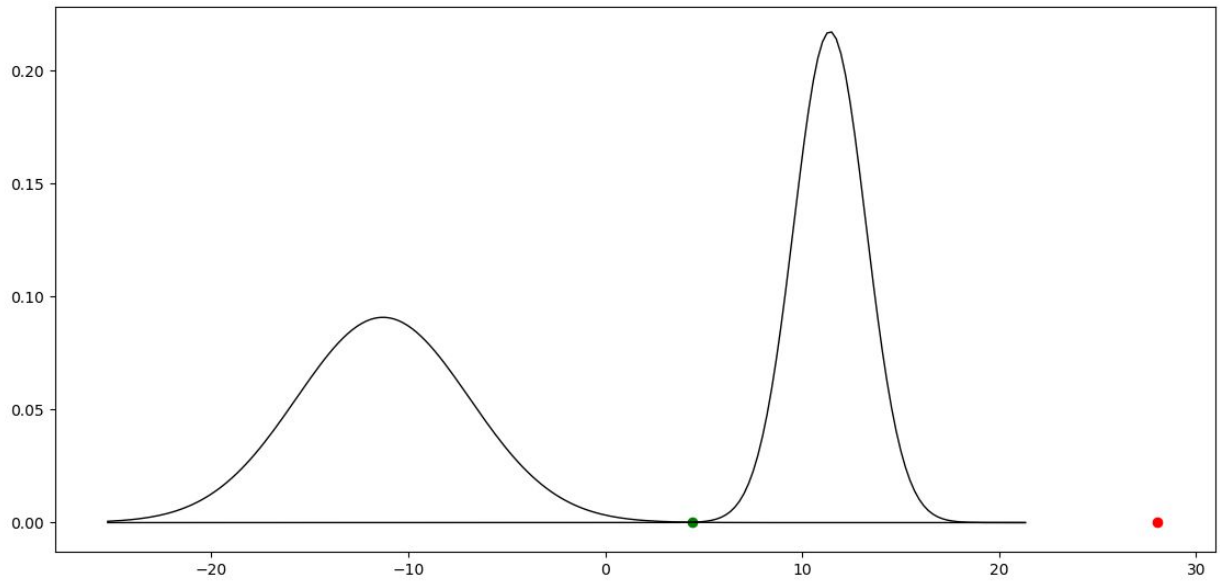
The figure shows the points after collapsing them on the classification vector, class0 points are in green and red points are the class 1 points. Which appear to be well separated

Fitting normal distribution to the transformed data points.



The mean of class 1 points is -11.29 and standard deviation is 4.39 , Class 0 has mean 11.41 and std deviation of 1.83.

Finding the intersection points



The point of intersection is calculated the same way as done before. The suitable discriminating point happens to be 4.41411.

So following the same process of classification on testing set and selecting 4.41411 as the discriminating point,the results were as follows

Results

df_confusion - DataFrame

Index	0	1
0	100	0
1	0	100

Accuracy of 100% and F score of 1.0 was obtained.

Summary

Dataset 1
Acc = **0.99**
F score = **0.99009.**

Dataset 2
Acc = **1.0**
F score = **1.0**

Conclusion

The high accuracy can be justified as the means of the two classes have been well separated i.e the difference of means is maximized and the covariance matrices are minimized.
In dataset 1 only 2 points are misclassified as a result.
In dataset 2 all the points are well separated hence 100% accuracy.