BITS F464 – Machine Learning

Assignment 1

Group Members:

Garvit Arora(2016B3A70462H)
Shivaank Agarwal(2016B4A70675H)
Parth Anand(2016B4A70873H)

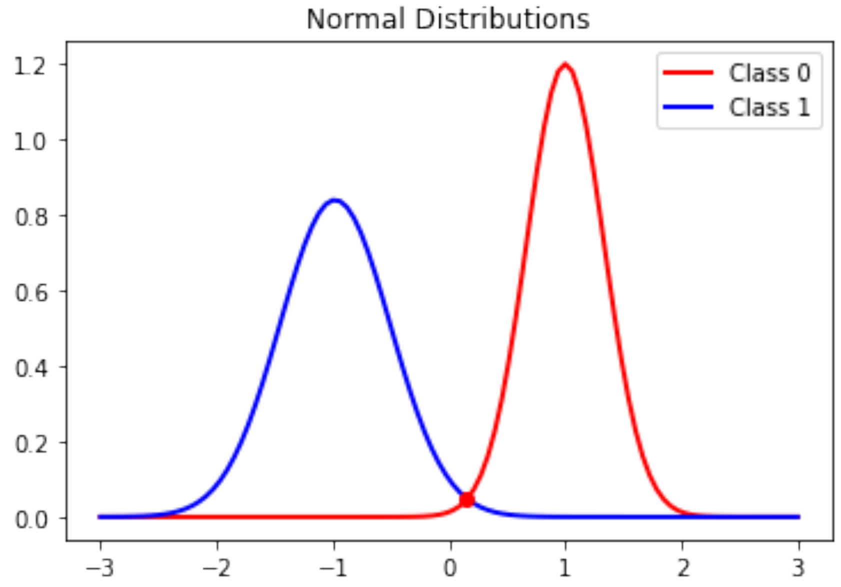
1) Fisher's Linear Discriminant Analysis:

Files included:

1. Fischer's LDA Dataset1.ipynb
2. Fischer's LDA Dataset2.ipynb

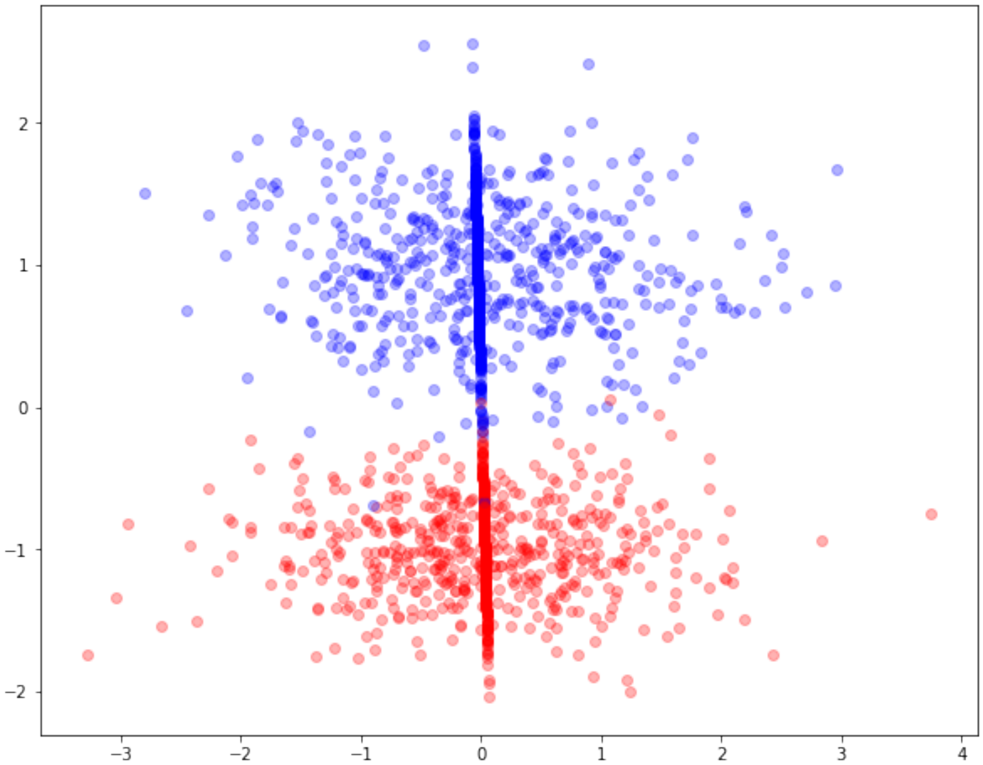
Fischer's Linear Discriminant Analysis on a1_d2.cv

Normal Distribution between Class 0 and Class 1



Separation point = 0.15151515151515138

Visualisation of points and separation of the two classes



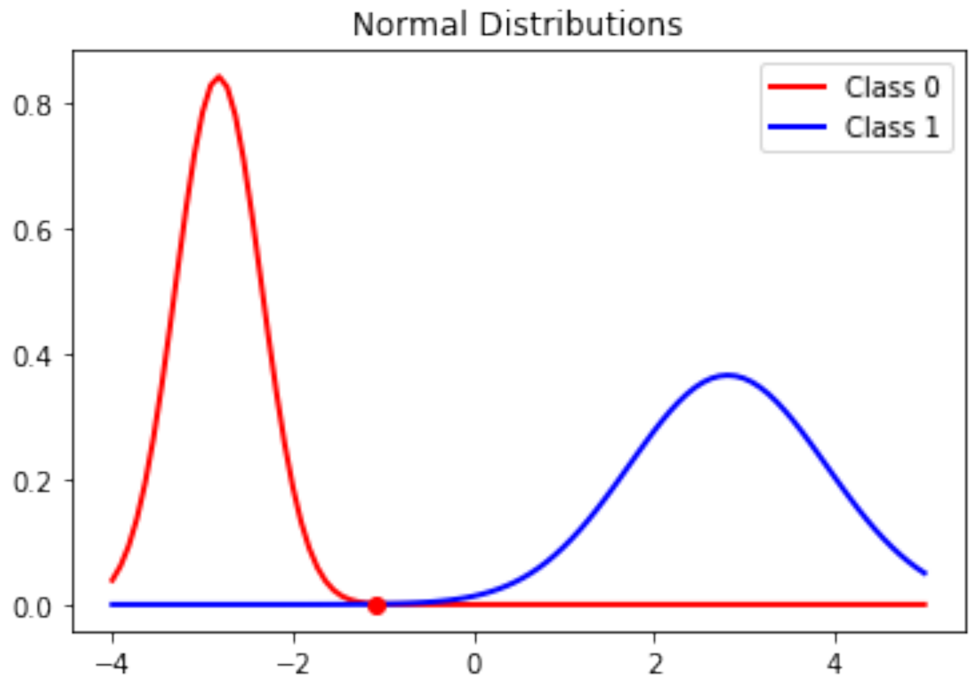
Results:

True positives **= 497**
True negatives = **498**
False positives = **3**
False negatives = **2**

Accuracy : 0.995
F1-Score: 0.994994994994995

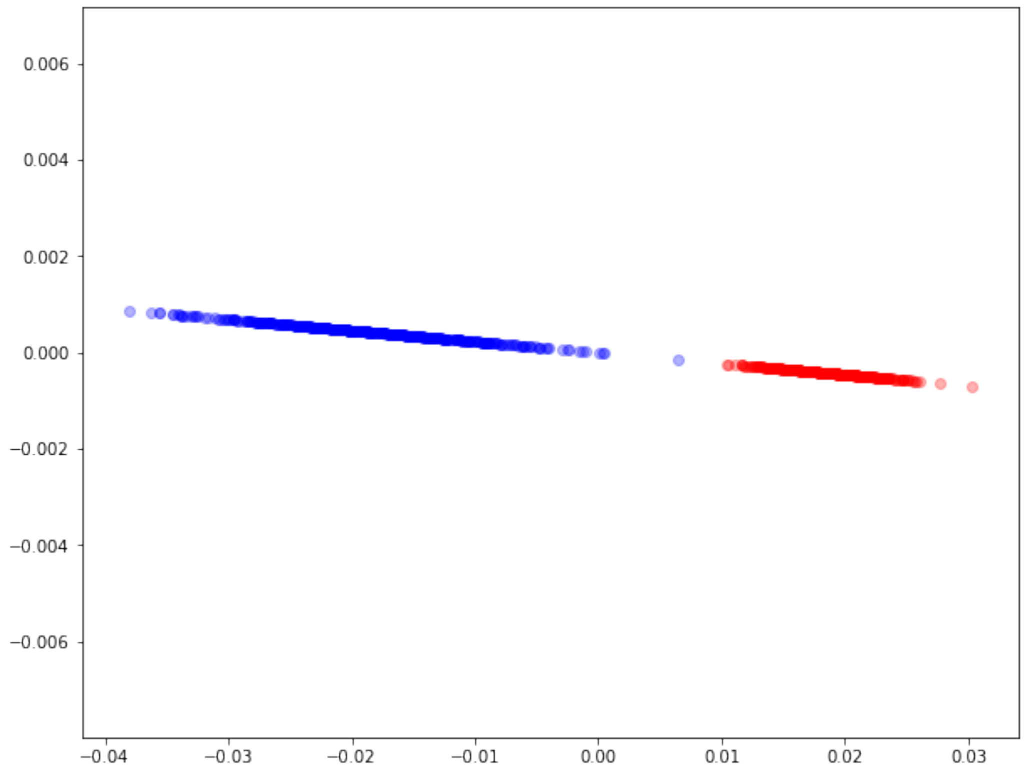
Fischer's Linear Discriminant Analysis on a1_d2.csv

Normal Distribution between Class 0 and Class 1



Separation point = -1.0909090909090908

Visualisation of points and separation of the two classes



Results:

True positives **= 500**
True negatives = **500**
False positives = **0**
False negatives = **0**

Accuracy : 1.0

F1-Score: 1.0

2) Naïve Bayes:

Files included:
Naïve Bayes.ipynb

Features: The words of the sentence are the features for each vector. Prior probabilities are the frequency of occurrences of positive and negative sentences in the training data.

Smoothening the frequencies of words did not yield improvements.
For a test sentence, if a word does not occur in positive(or negative) sentences even once, it's frequency is considered to be one. This may be considered as artificial smoothening.

Pre-processing: The dataset was pre-processed to improve the model performance and to reduce the errors due to the raw data. The entire data text was converted to lowercase to avoid difference between capitalised words and other words. Also, all punctuation marks(, . ! etc) were removed so that only words were considered in calculating the conditional probability.

Accuracy:

5-fold cross validation was used for evaluating the model's performance:

The accuracies for each iteration were:

```
1) 0.765
2) 0.815
3) 0.83
4) 0.8
5) 0.855
```

The F-scores for each iteration were:

```
1) 0.7344632768361581
2) 0.8229665071770335
3) 0.8282828282828283
4) 0.7999999999999999
5) 0.8449197860962566
```

Mean ± standard dev = **0.813 ± 0.03009983388658481**