BITS F464

Machine Learning

Assignment - 1 Final Report

Submitted by:

Pranjal Gupta (2017A7PS0124H)
Ujjwal Raizada (2017A7PS1398H)
Simran Sandhu (2017A7PS1454H)

Fisher's Linear Discriminant Analysis

Dataset

Two datasets were given a1_d1.csv containing two features and a1_d2.csv containing 3 features

Fisher's Algorithm

Maximize difference of two means and minimize $s_1^2+s_2^2$
Maximize $(w^T(m_1-m_2))^2/ (s_1^2+s_2^2)$
Result: $S_W w \propto m_1 - m_2$
Where w is the projection vector
m_1 is mean of class 1 points after being collapsed
m_2 is mean of class 2 points after being collapsed
s_1 is variance of class 1 points after being collapsed
s_2 is variance of class 2 points after being collapsed,
S_W is covariance matrix

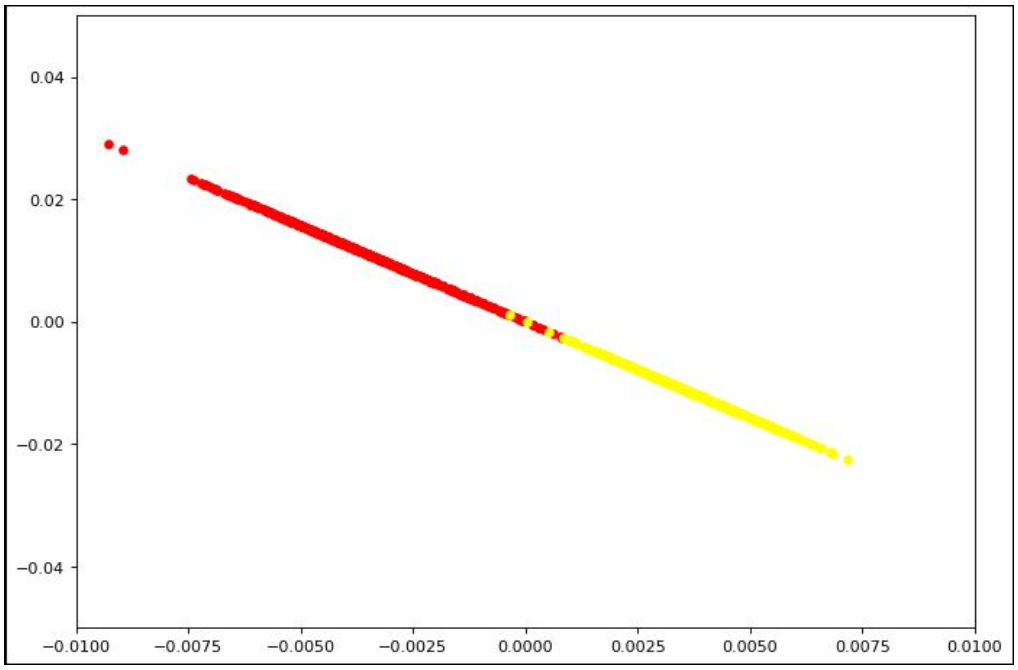
Steps

1) First the projection vector direction is found
   Projection vector direction=(covariance inverse)(mean1-mean2)
2) Next step is to find projection of points on the Projection vector
3) Next is to plot their normal distribution and find the intersection point
4) Further find the projection of points on the vector and if their value is greater than x coordinate of threshold point, classify it as class 0 else classify it as class 1

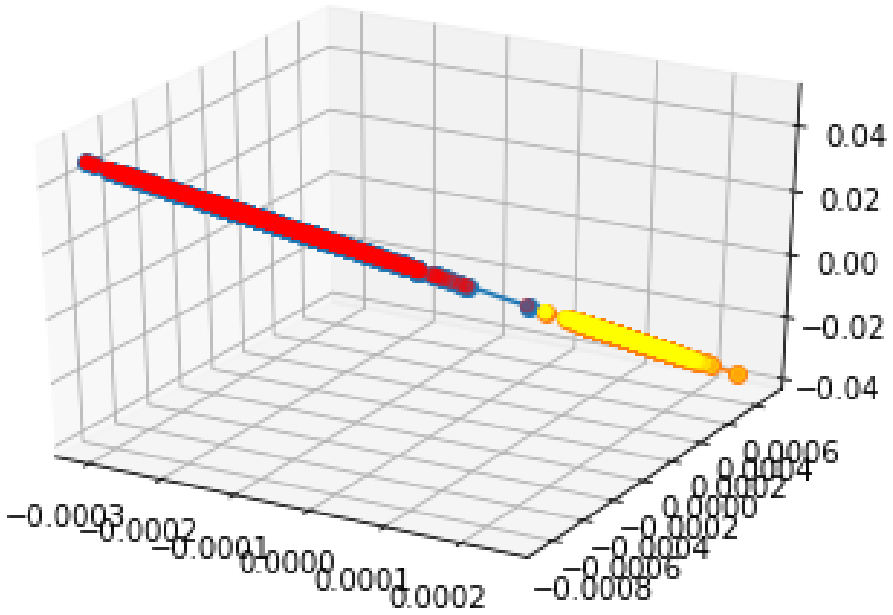
Results

Points in dataset 1 are classified with accuracy 99.29% and F_Score 0.99 and points in dataset 2 are classified with accuracy 100 % and F_Score 1

Visualizations



Visualization for dataset 1



Visualization for dataset 2

Naive Bayes

Preprocessing

The dataset used for this assignment consists of a customer's review followed by his sentiment value (0 or 1) separated by a tab space. Without any of the preprocessing steps the accuracy obtained is ***72.2± 2.20%***. The following are the preprocessing steps and the accuracies obtained with them:

1) Convert all upper case characters to lower case characters: ***75.1±4.02%***
2) Remove all numbers ***75.2±3.96%***
3) Remove all punctuation ***76.9±1.93%***

Dataset Preparation

Our dataset is constructed by using a python dictionary which contains the information about occurrence of each word in each sentiment category. Format for the dataset is

Dataset = { 'word1' : { '0' : , '1' : } , 'word2' : { '0' : , '1' : } ....}

Steps

For each testing point its test probability for that category multiplied with category probability is compared for the two categories. i.e. Compare:

P(A|B) = P(B|A) * P(A) / P(B)

A : Probability of being Category 0 or 1

B : Test data

P(A|B) : Category given the test data

P(B|A) : Test data given the category

We can ignore P(B) in the denominator since it doesn't change

The steps to calculate test probability for each category is as follows:

1) Each out of vocabulary word is assigned the probability 0.5

2) For in-vocabulary word its probability is calculated in a weighted fashion as follows:
   basic_prob=(no_of_occurences_in_category)/(total_no_of_words_in_category)
   P(B|A)=0.5+total_no_of_occurences*basic_prob/1+total_no_of_occurences

3) 5-fold cross-validation is performed on the dataset to determine the mean accuracy and mean F-score.

Conclusion

Converting all upper case characters to lower case characters resulted in significant improvements (by 3%) to accuracy.