Out[1]: Toggle Code

Problem 1

Methods

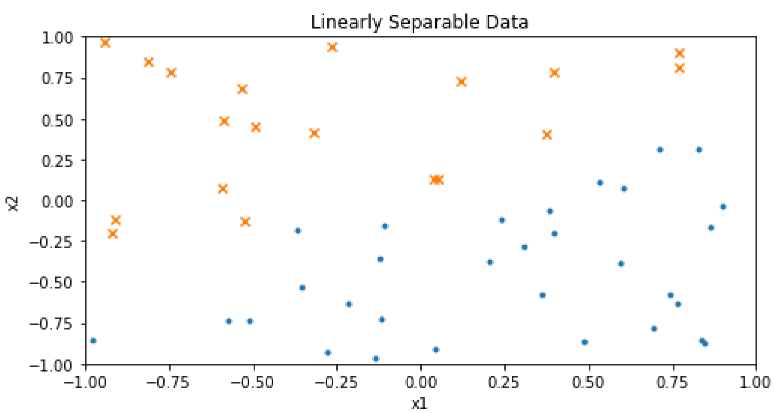
In this problem, we examine the LDA and PCA methods on various datasets. We refer to the notes on lecture 11 and 12 on the procedure for each of these algorithms.

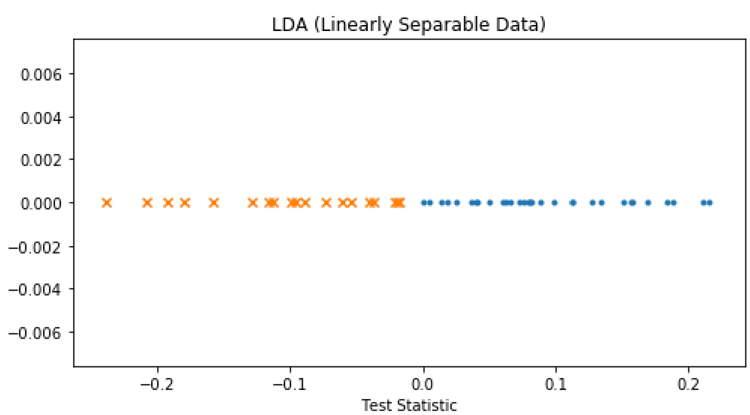
The two methods were tested on 3 datasets. The 1st is a dataset that is known to be linearly separable. This was accomplised by generating a random weight vector, applying it to randomly generated data, and creating labels by arbitrarily setting the dection boundary to approximately spit the data in half.

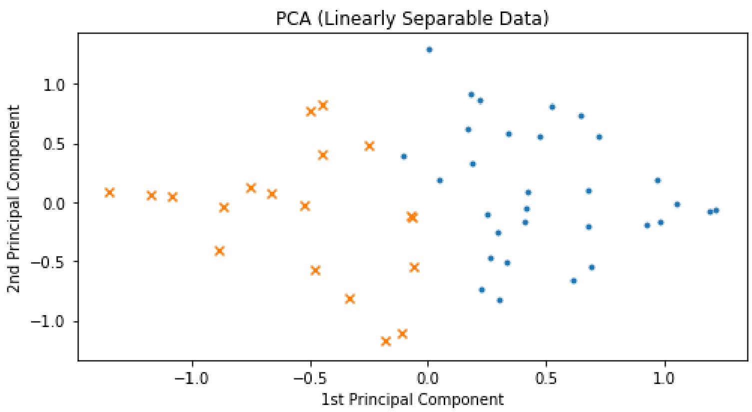
The 2nd dataset was generated similarly to the 1st, with exception of squaring the data, before applying the weight vectors. This forces the data to be non-linearly separable.

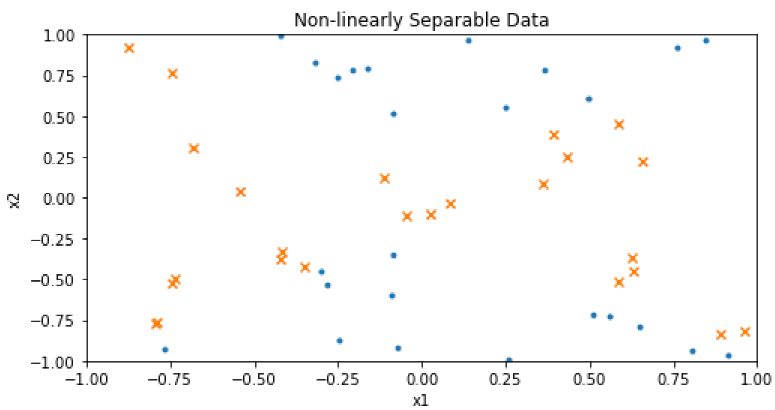
The 3rd dataset are two 2d uniform gaussian distributions set to different means and standard deviations.

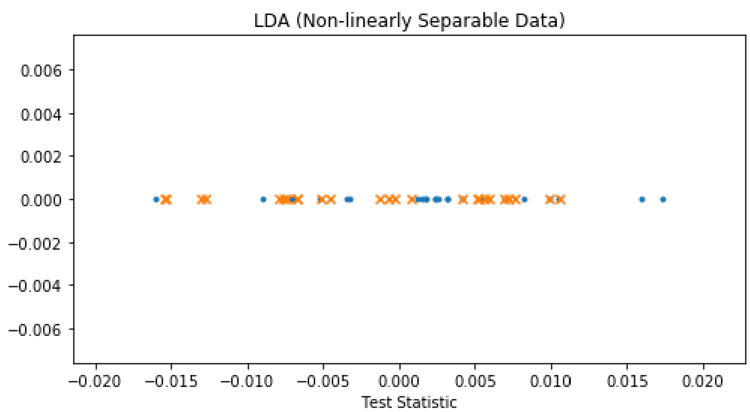
Results

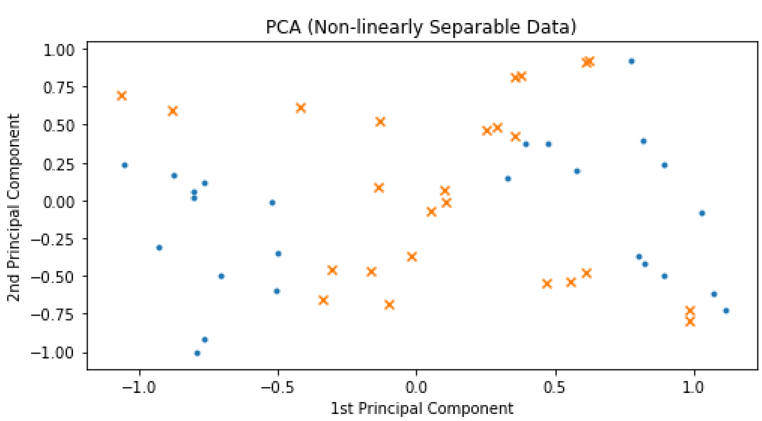


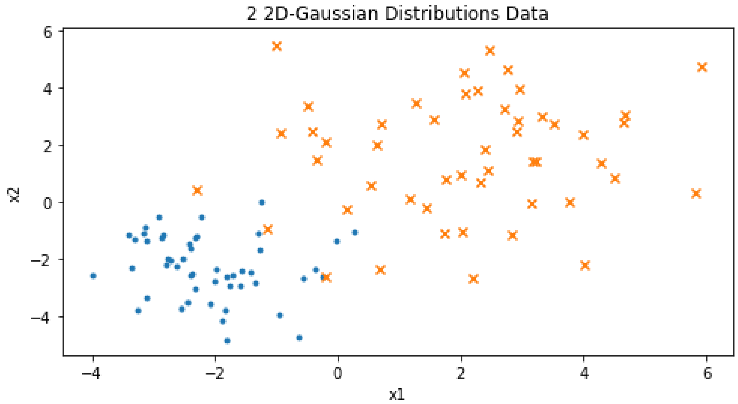


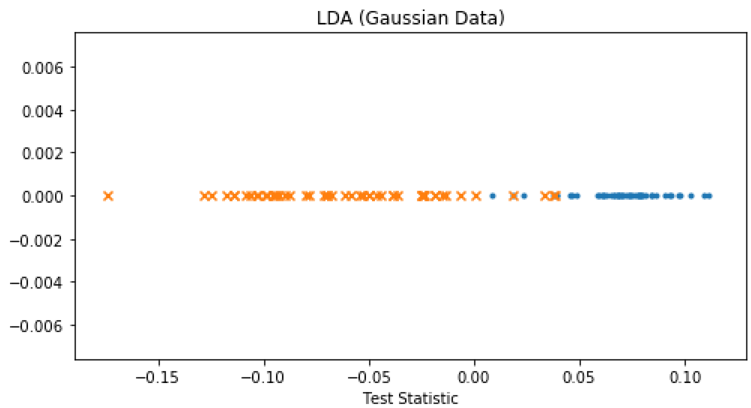

PCA (Linearly Separable Data)
2nd Principal Component
1st Principal Component
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

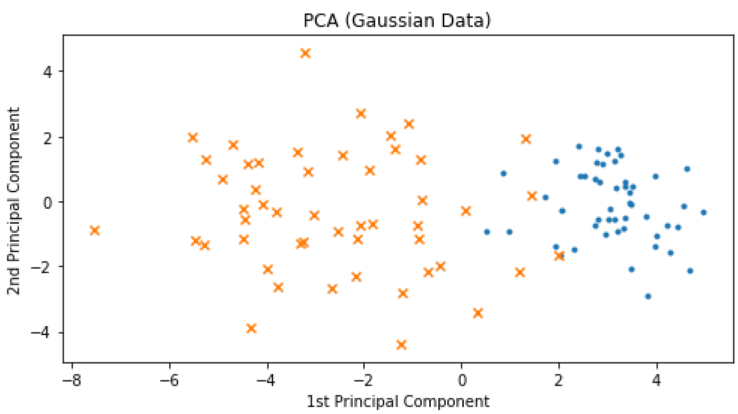

Non-linearly Separable Data
x2
x1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


LDA (Non-linearly Separable Data)
Test Statistic
0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
−0.020
−0.015
−0.010
−0.005
0.000
0.005
0.010
0.015
0.020


PCA (Non-linearly Separable Data)
2nd Principal Component
1st Principal Component
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0







Discussion

For both methods, we can see that they are able to cleanly separate linearly separable data well. Both are linear methods and so this would be expected. Counter to this, they both fail at separating non-linearly separable data.

The generated gaussian distribution dataset shows the difference between the LDA and PCA methods. We see that the LDA method can separate the two groups more cleanly than the PCA method. This is due to the fact that the LDA projects the data to a subspace that maximizes the separation between the two classes, while PCA merely finds the directions of maximal variance. Thus LDA can provide cleaner separations of classes over PCA. However, we should note that the LDA method requires class labeling, while PCA does not.