Avik Bag

Professor Matt Burlick

CS 383 – Assignment 1

25th January 2017

# Theory Questions

Question 1

1. The goal is to perform Principal Component Analysis on the raw data that is provided.

This raw dataset will need to be standardized. This is done by finding the average and standard deviation along the columns. Using these values, the data is centered by subtracting the average value and then dividing the standard deviation along the respective columns. Here are the values computed below.

After standardization:

Now that the data is centered, the covariance matrix can be computed using the following equation.

Therefore, the computed covariance matrix is the following

The next step is to evaluate the eigenvalues and eigenvectors based on the covariance matrix. The computation process is as follows;

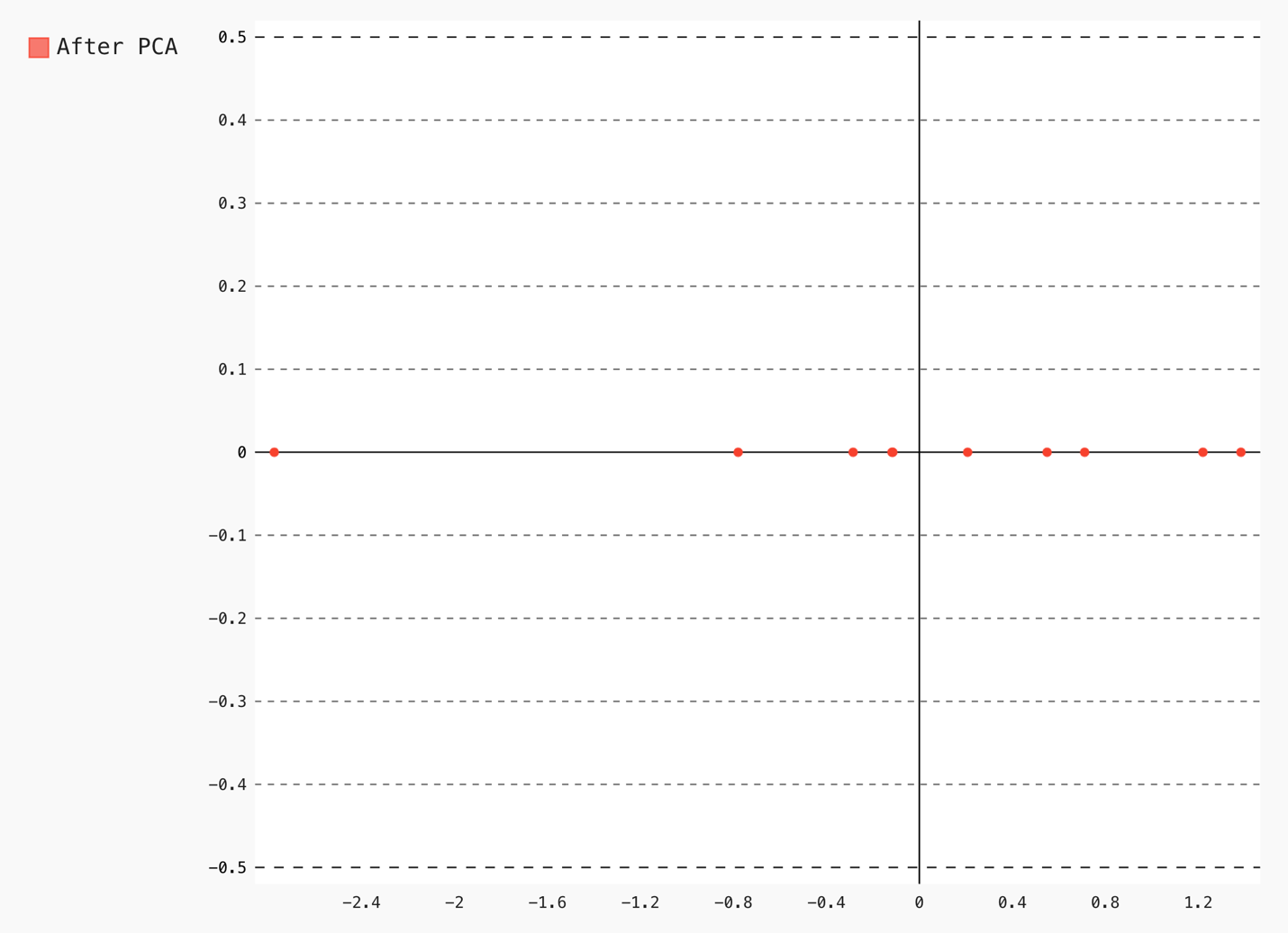
Using Quadratic formula;

Since we are interested in the best eigenvector to project the data on, the the eigenvector corresponding to the highest eigenvalue will be used. This is computed with the help of python’s numpy library.

Let = 1.408

Finally, the standardized data is projected onto the eigenvector to complete the PCA process.

1. On Projecting the data on the principal component, this is the result.



Question 2

1. The goal here is to find the information gain from the two classes. For this question, the assumption will be made that column 0 will correspond to pi, and column 1 will correspond to ni.

For Feature 1;

|  |  |  |
| --- | --- | --- |
| K | pi | ni |
| -8 | 1 | 0 |
| -5 | 1 | 0 |
| -4 | 0 | 1 |
| -3 | 1 | 0 |
| -2 | 1 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 2 |
| 3 | 0 | 1 |
| 11 | 0 | 1 |

For Feature 2;

|  |  |  |
| --- | --- | --- |
| K | pi | ni |
| -3 | 0 | 1 |
| -2 | 1 | 0 |
| -1 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 0 |

1. Thus here we see that the information gain for feature 1 is 1.4, as opposed to a value of 0.4 for the information gain for feature 2. Based on this information, feature 1 is more discriminating.
2. The next step is to use the Linear Discriminant Analysis for feature projection.

The first step is to concatenate the entire dataset and then standardize. The standardized data is as follows.

We then split it to get the respective class data

Table 1: Standardized data with mean and std

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| Class data |  |  |
|  |  |  |
|  |  |  |

Calculating Scatter Matrices is done by using the following equation

These are the computed scatter matrices.

|  |  |  |
| --- | --- | --- |
| Class data | 1 | 2 |
| Scatter Matrix |  |  |

Within class scatter matrix

Between class scatter matrix

Eigen decomposition is performed on . This is done with the help of python.

Eigenvalues = [0.1459, 0.3473]

Eigenvectors =

Since the value at index 1 for eigenvalues is the greatest, the corresponding eigenvector that will be used is column 1 of eigenvector, i.e, .