Introduction to Data Science - Assignment 4

Shahriyar Mahdi Robbani

March 10, 2020

Exercise 1

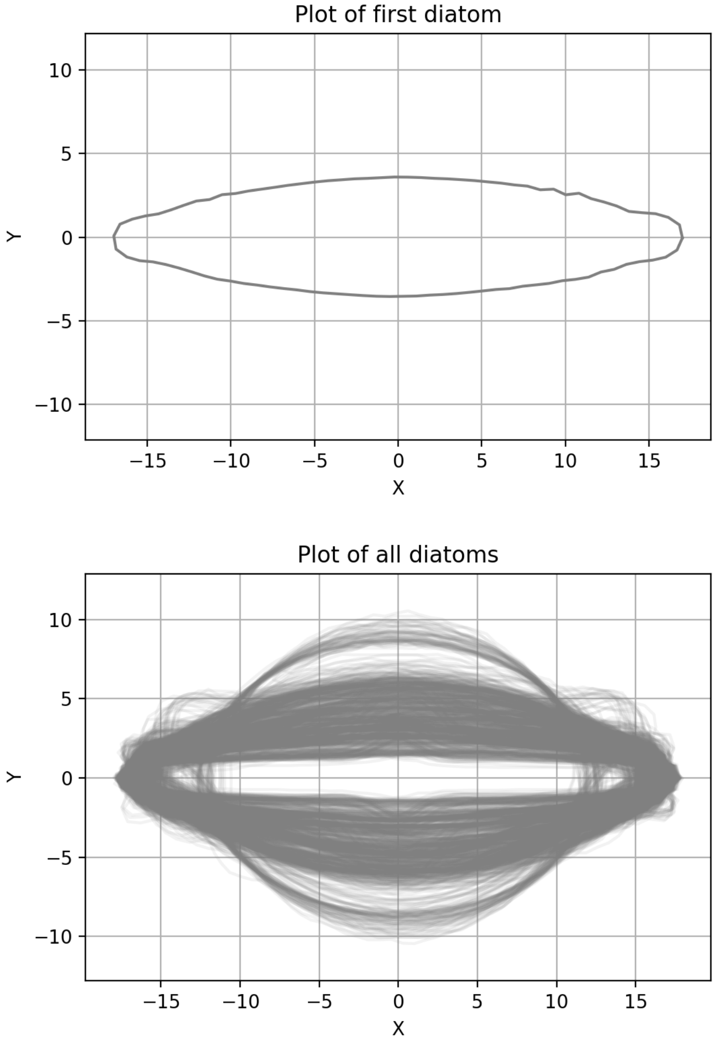

Figure 1: Plot of diatoms

The diatom dataset seems to consist of two distinct types of diatoms. The majority of the diatoms are thin and elongated, seen from the dark lines similar to the plot of the first diatom. The remaining diatoms are shorter and more round which can be seen from the more faint circular outline.

Exercise 2

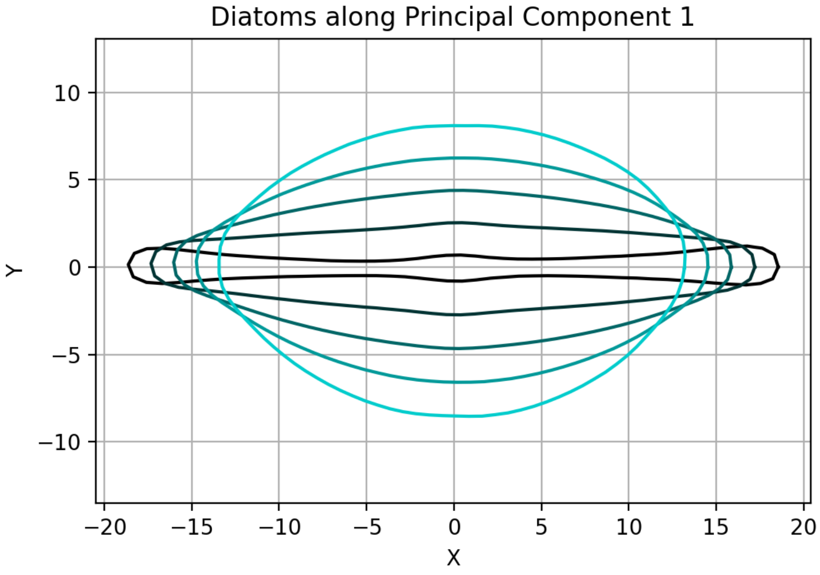

Figure 2: Plot of diatoms along Principal Component 1

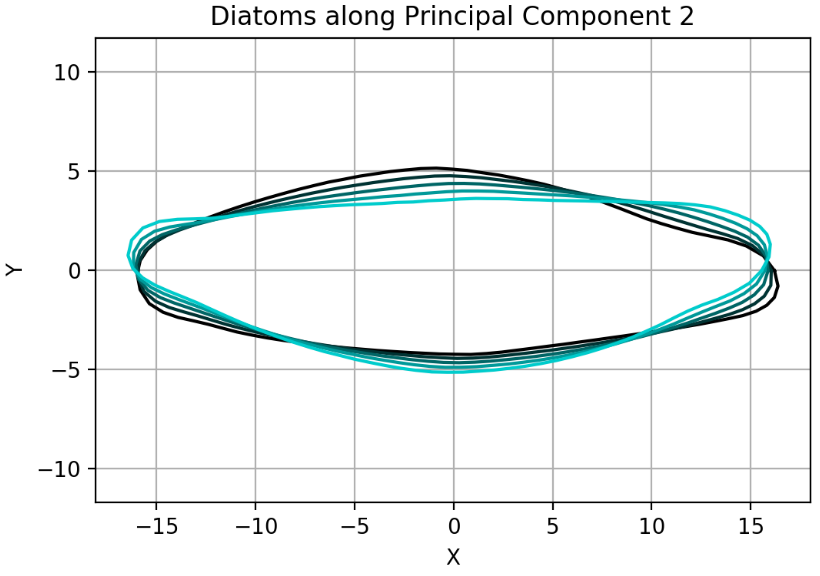

Figure 3: Plot of diatoms along Principal Component 2

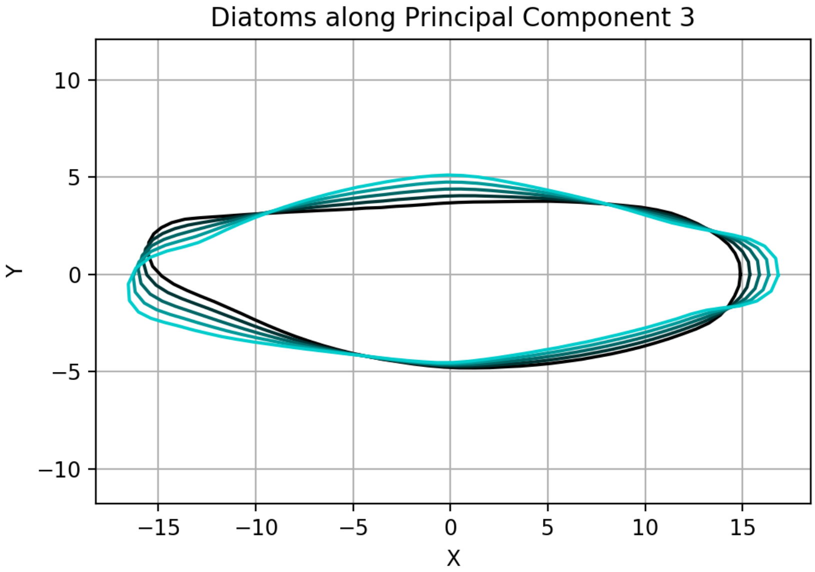

Figure 4: Plot of diatoms along Principcal Component 3

Principal Component 1 captures the two different shapes present in the diatom dataset.
Principal Component 2 captures slight differences in the orientation of the thinner diatoms.
Principal Component 3 also captures slight differences in the orientation of the thinner diatoms.

Exercise 3

1. **Centering**
   Centering data before performing PCA causes all the data points in the PCA plot to be centered around the origin. This does not change the eigenvectors or eigenvalues so it is not necessary to perform. However, centering data is considered good practice.

2. **Standardization**
   Standardizing data before performing PCA reduces the variance of the data, which results in that component of the data to have a smaller effect on the eigenvectors. If the data consists of highly varying, unrelated units of measurement then it is a good idea to standardize data. If all the units of the data are the same then standardization is not necessary.

3. **Whitening**
   Whitening data converts its covariance matrix into the identity matrix. This means all covariance information is lost so PCA will not work on this data. Therefore data should be whitened before PCA.

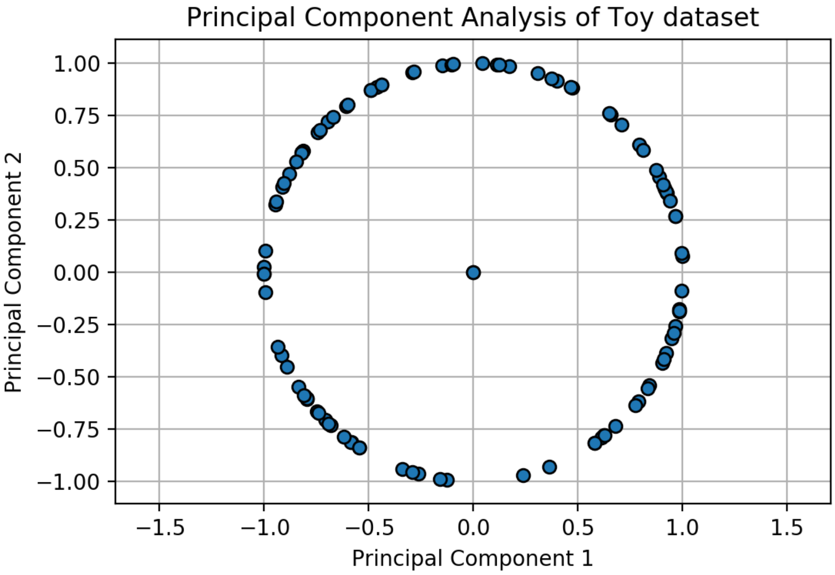

Figure 5: Principal Component Analysis of Toy dataset

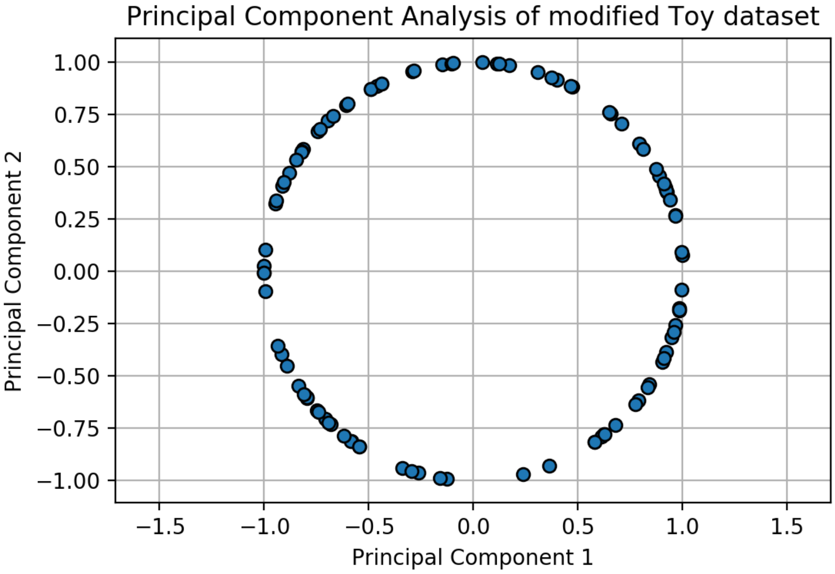

Figure 6: Principal Component Analysis of modified Toy dataset

Removing the last two data points causes the points in the center of the PCA plot to disappear,

Exercise 4

I used the KMeans function I created in the previous assignment to The cluster the data. That function calculates a list of distances between each data point and the centroid for k centroids, and converts these lists into a matrix. It then finds which data point has the smallest distance to a centroid, and then assigns that data point to a cluster. This results in k clusters. The new centroids are then calculated by finding the mean position of each cluster. The loss is defined as the sum of squared distances between each point and the corresponding centroids and is used as a measure of how good the centroids are. The algorithm continues looping until the loss converges, indicating no change in the centroids.

I used the MDS function created in the previous assignment to project all the data points to Principal components 1 and 2. In order to color each of these transformed points, I checked the column of labels provided with the data. Since each row of the data has a corresponding label, I used that label to subset the data into crops and weed, and colored each subset differently. The centroids were transformed by calculating the eigenvectors of all the data points, and then using the first two eigenvectors to transform the centroid values calculated from the K Means Clustering function.

```
Centroid 1 Projection: [-1596.80176436   123.75257989]
Centroid 2 Projection: [1404.7053115  -108.86505148]
```

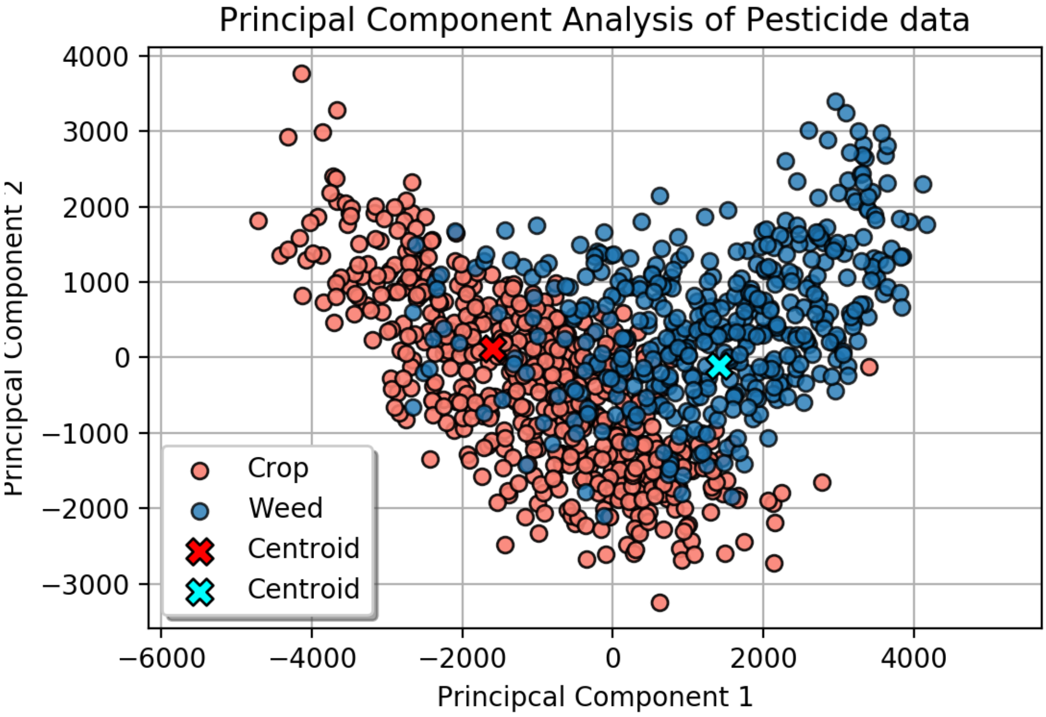

Figure 7: Principal Component Analysis of Pesticide data

From the diagram it can be seen that there are many blue points close to the red centroid, and a few red points close to the blue centroid. This means the K means clustering algorithm incorrectly classified these points, since the algorithm would just assign them to the nearest centroid. This means the K means clustering algorithm is too simple to find the true underlying distinction between the weed and crop data so a more sophisticated algorithm is required to cluster the data.