Report Introduction to Data Science – Assignment 4

Exercise 1

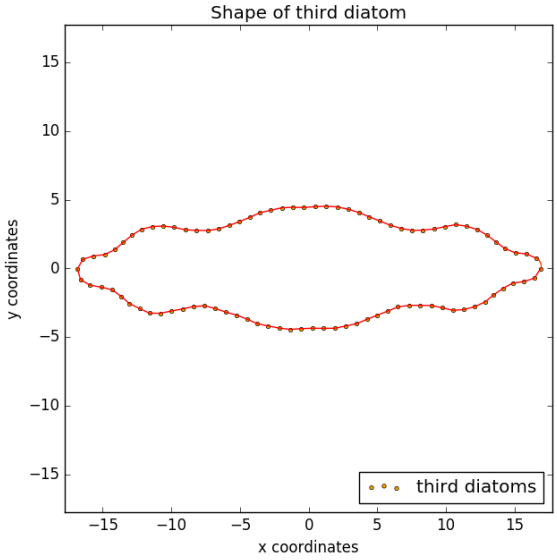

Figure 2 Shape of the third diatom

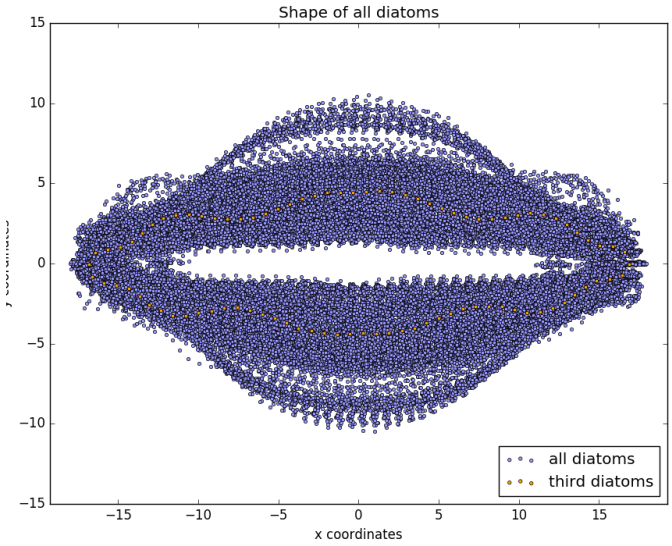

Figure 1 Shape of all diatoms

In Figure 1 are plotted the points that describe the third diatoms shapes from the dataset provided. Each diatom is described with 90 points and their x and y coordinates. The points are interpolated with the red line.
In Figure 2 are plotted the shapes of every diatoms from the dataset. The orange points are the ones from plot in figure 1.
From this plot underlines a considerable enlargement of the shape of the diatoms along the y axis and a smaller variances along the x axis. Furthermore, some variations on the externals upper sides of the cells are registered.

Exercise 2

The plots underneath describe the spatial variance of five cells and plotting some instances of the first three principal components. The cells plotted follows the pattern below:

$plot1:\ m-2\sigma_1 e_1\ \ m-\sigma_1 e_1\ \ m\ \ m+\sigma_1 e_1\ \ m+2\sigma_1 e_1$
$plot2:\ m-2\sigma_2 e_2\ \ m-\sigma_2 e_2\ \ m\ \ m+\sigma_2 e_2\ \ m+2\sigma_2 e_2$
$plot3:\ m-2\sigma_3 e_3\ \ m-\sigma_3 e_3\ \ m\ \ m+\sigma_3 e_3\ \ m+2\sigma_3 e_3$

Where e_i are the eigenvectors and σ_i are the standard deviations of the data projected on the PCs.

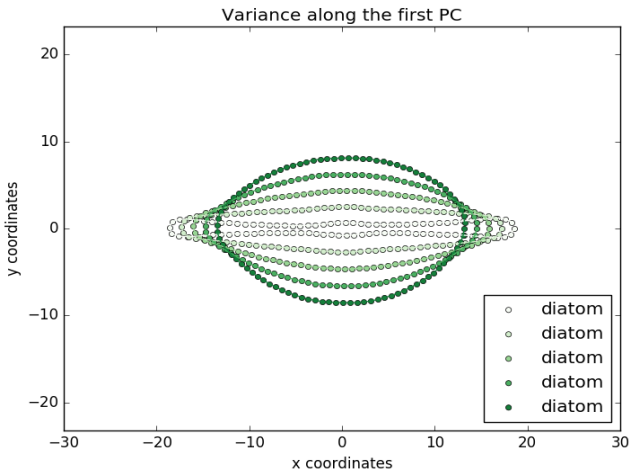

Figure 3 Variance along first principal component

In figure 3 is plotted the variance along the first principal component. The plot highlight a major variance change along the y axis. So the first principal component shows a variation in what could be called a 'swelling' of the cells. This swelling has also an effect on the elongation of the cells. In fact, the variation along axis x shows a difference between the white diatom, more elongated and the dark green one, more 'swollen'.

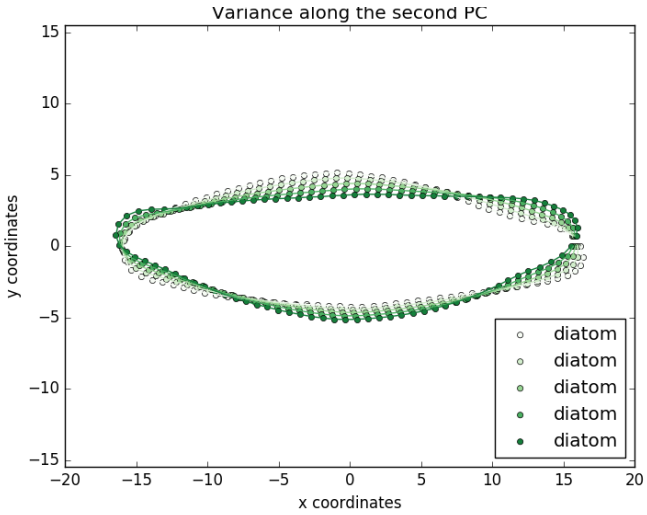

Figure 4 Variance along second principal component

In figure 4 is plotted the variance along the second principal component. From the plot, the most important variation could be considered as a torsion of the shape between the cells along the horizontal axis. Moreover, the variance between the white diatom and the dark green diatom along axis y is not as relevant as in figure 3. Moreover, comparing the variations between figure 3 and 4 along the x axis, the one in figure 4 is almost null.

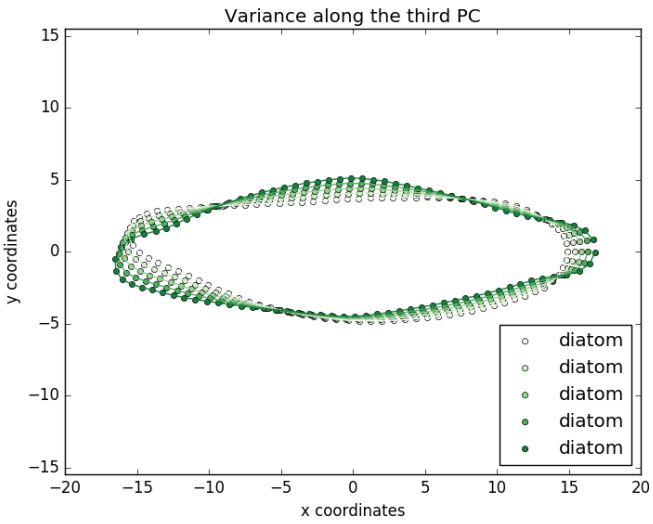

Figure 5 Variance along third principal component

In figure 5 is plotted the variance along the third principal component. As in figure 4, the variance describes a twist but on the other direction. Nevertheless the variation along axis x is a bit more relevant than in figure 4 .

Exercise 3

a)

Centering: Computing PCA results from the covariance matrix of a data set, meaning that the principal components are the eigenvectors of the covariance matrix that correspond to the largest eigenvalues, then centering the data doesn't have any effect on the PCA results. This is explained because the covariance is $\sigma_{xy} = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})$ being x and y two variables of a data set. If we center the variables: $x' = x - \bar{x}\ and\ y' = y - \bar{y}$ where x' and y' are the centered variables. When computing the covariance $\sigma_{xy} = \frac{1}{n-1}\sum_{i=0}^{n}(x'_i - \bar{x}')(y'_i - \bar{y}')$ but because our variables are centered then the mean is equal to zero. This means that the covariance matrices are the same. Nonetheless, it is always useful to center the data to remove any bias in the inputs by translating the origin of the system.

Standardization: The standardization process is used to make sure that all the features in the data set have the same scale. This is done by dividing all the data points by their standard deviation. This affects the covariance matrix from which the PCA is performed and thus the results of the PCA. It is a good idea to standardize the data otherwise the 'out-of-scale' points would affect the variance and dominate the outcome.

Whitening: The whitening transformation is multivariate data set which results in a decorrelated data set that has uniform variances on all diagonals. This means that using a decorrelated data set to perform PCA it will for sure affect the results as PCA is based on the variances of each points. It is a good idea though to perform whitening after PCA.

b) The plot in figure 6 shows the points from the toy dataset projected onto the first two principal components.
In figure 7 are plotted the points from the same dataset but keeping out the last two points.

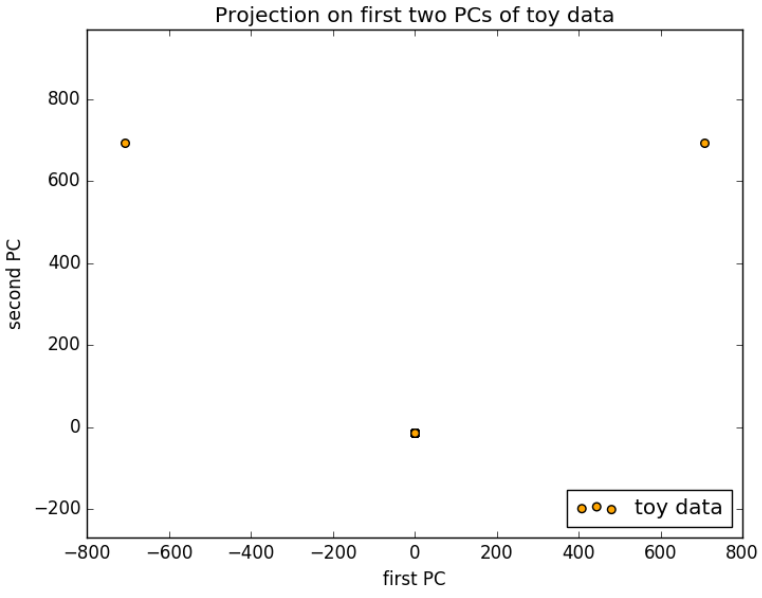

Figure 6 projection on the first 2 PCs

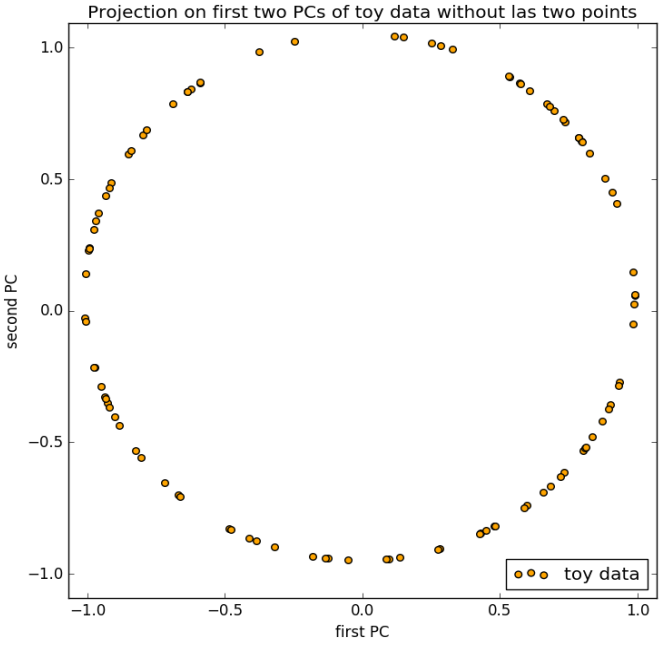

Figure 7 projection on first 2 PCs leaving out last two points

If we take a look at the last four points of the dataset:
`−0.902316 −0.431076 −0 −0 |−0.944341 0.328969 −0 0 | 0 0 1000 0| 0 0 0 1000` it is noticeable that the last two points have very different values compared to the other. When we perform PCA on the entire dataset, the last two points have a larger variance that will

dominate in the outcome. Taking those two points into considerations leads to have the plot on figure 6 where the variance on the first two PCs is described by only three points.
On the other end, if we remove the last two points then it could be said that the distribution of the sample is more uniform and the projection of the points over the first two PCs results in the plot in figure 7 showing the hidden structure of a circle.

Exercise 4

The plot in figure 8 shows the data points from the weed and crop train data set projected onto the first two principal components and coloured according to their original labels. The green diamonds are the centroids of the clusters found by applying k-means algorithm.
In the file **clustering2.py** PCA is performed on the dataset. The first two eigenvectors found are then used to reduce the dimensionality of the data set. The data points are plotted in figure 8 assigning the colours to the points belonging to the weed class and to the crop class.

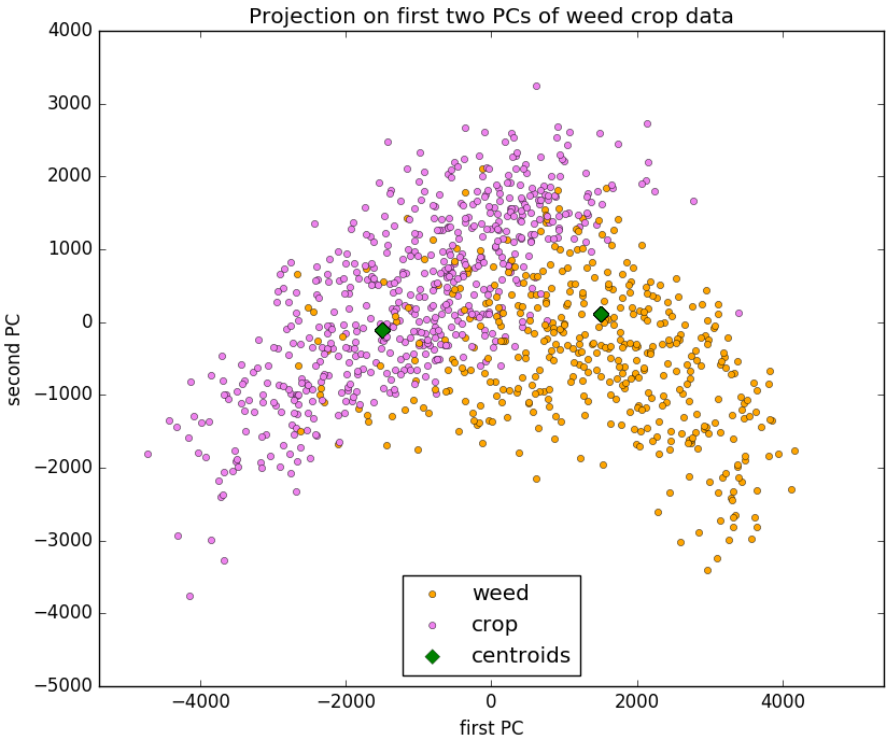

Figure 8 projection on first 2 PCs of weed and crop data points

The k-means algorithm is performed in the function **k_means()** that takes as arguments the data set, the number of cluster to look for, two in this case, and the number of maximum iteration, 300 for this example. The algorithm takes the first 2 points in the data set as starting centroids, it computes the distance between each point in the data set and the two centroids and append the index of the minimum distance in a list. The average value the is the used to redefine the centroid. A dictionary with the number of centroids (2) as item and its features as values is then returned.
The centroids are then reduced in dimensionality using the eigenvectors from PCA on the entire dataset and their projections are plotted in figure 8 as green points.
The centroids seems to be in the middle of the point clouds so it can be said that the clusters found are meaningful.