Introduction to Machine Learning
Assignment 3

Group 31
Stijn Kammer (s4986296) & Ramon Kits (s5440769)

October 5, 2022

1. Introduction

The method used in this assignment is the use of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN is a widely used data clustering algorithm which is commonly used in data mining and machine learning. This algorithm is a density based clustering algorithm. It is a density based clustering algorithm because it uses the density of points in a region to determine if they belong to the same cluster. The algorithm is based on two parameters, the minimum number of points in a region and the maximum distance between points in a region.

2. Methods

DBSCAN groups together points that are close to each other based on two parameters: ϵ (epsilon) and $minPts$. ϵ is the maximum distance between two points to be considered close to each other. The $minPts$ parameter is the minimum number of points that need to be close to a point for it to be considered a core point. Every point that is not a core point but is inside the ϵ of a core point is considered a border point. All connected core points and border points form a cluster. Every point that is neither a core point nor a border point is considered noise.

In six steps, the algorithm works as follows:

1. For each point in the dataset, determine the number of points in a region around it.
2. If the number of points in the region is larger or equal to the minimum number of points, the point is a core point.
3. If the number of points in the region is smaller than the minimum number of points and it belongs to the neighbors of a core point, it becomes a border point.
4. If the number of points in the region is smaller than the minimum number of points and it does not belong to the neighbors of a core point, it becomes a noise point.
5. For each core point, determine the cluster it belongs to.
6. For each border point, determine the cluster it belongs to.

This algorithm has a few advantages. you can use it to find clusters of any shape and size. There is no need to specify the number of clusters beforehand. And is also very robust to noise and outliers. This algorithm works best when you choose the minimum number of points based on the

size and noise of the dataset. For two-dimensional data, the use of 4 as minimum number of points is recommended. for higher dimensional data, the use of two times the dimensionality of the data is recommended.

3. Experimental results

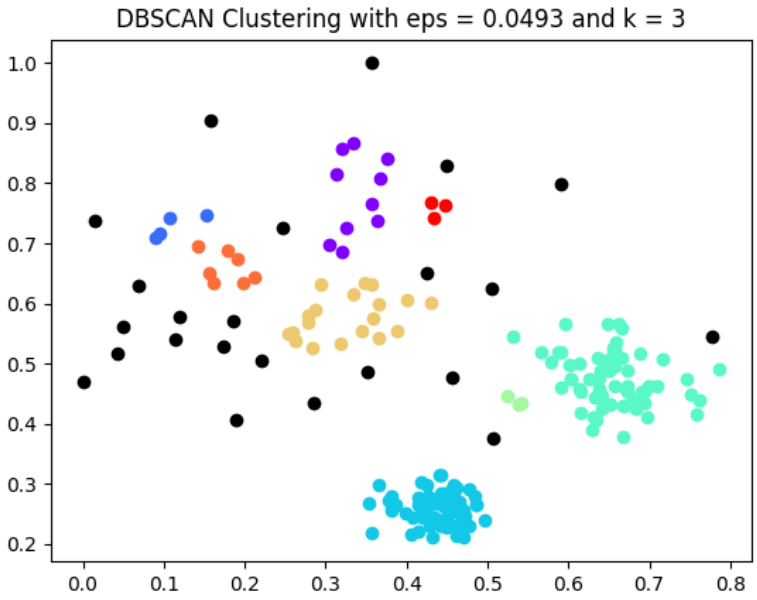

Figure 1: *DBSCAN with* $\epsilon = 0.0493$ *and* $k = 3$

Figure 1 shows the result of the DBSCAN algorithm with $\epsilon = 0.0493$ and $k = 3$. The algorithm found multiple smaller clusters and multiple noise points. The clusters are very close to each other and are not clearly visible in the figure.

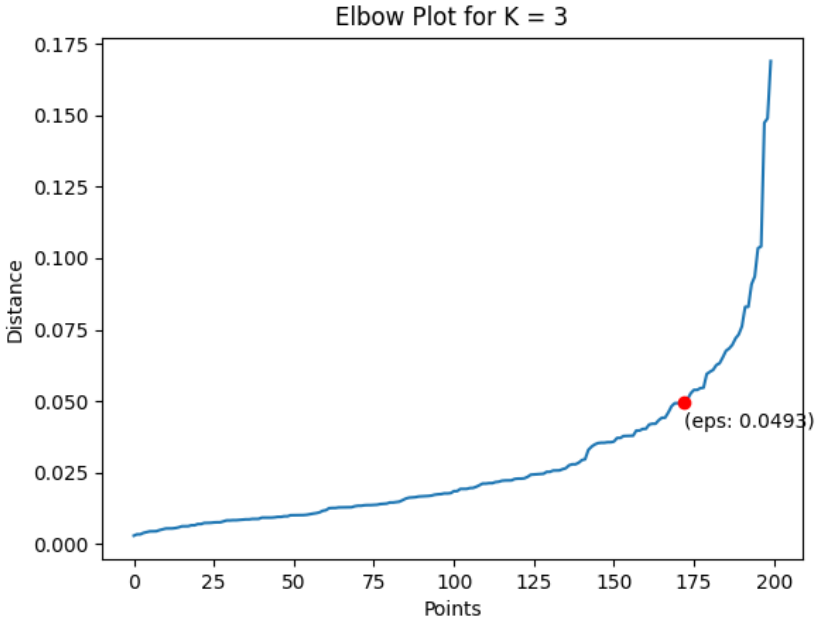

Figure 2: *Elbow plot with* $k = 3$

Figure 2 shows the elbow plot with $k = 3$. The red dot is the elbow point determined by the researchers. This should be the place with the greatest change in the slope of the curve.

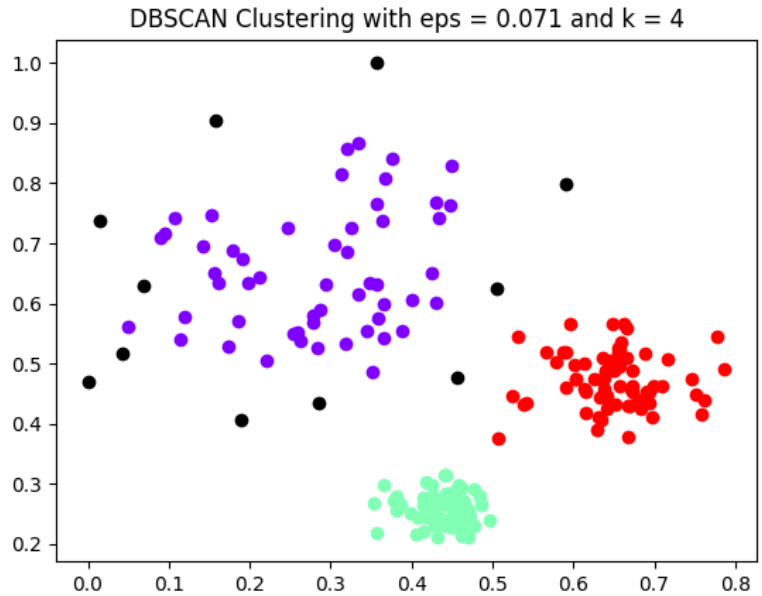

Figure 3: *DBSCAN with* $\epsilon = 0.071$ *and* $k = 4$

Figure 3 shows the result of the DBSCAN algorithm with $\epsilon = 0.071$ and $k = 4$. The algorithm found 3 clusters and multiple noise points. The clusters are clearly visible in the figure. The

clusters are not perfectly circular, but they are still clearly visible. The noise points are also clearly visible in the figure.

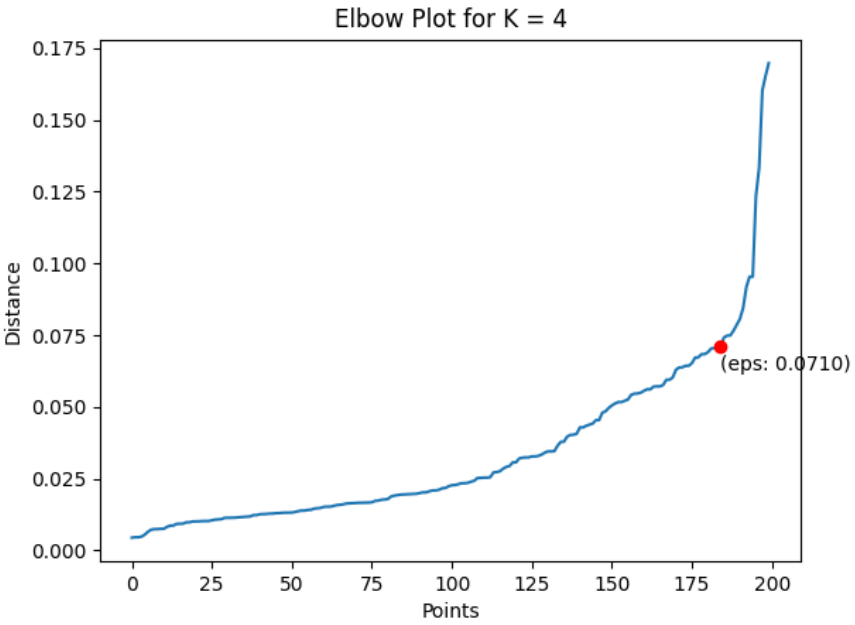

Figure 4: *Elbow plot with* $k = 4$

Figure 4 shows the elbow plot with $k = 4$. The red dot is the elbow point determined by the researchers. This should be the place with the greatest change in the slope of the curve.

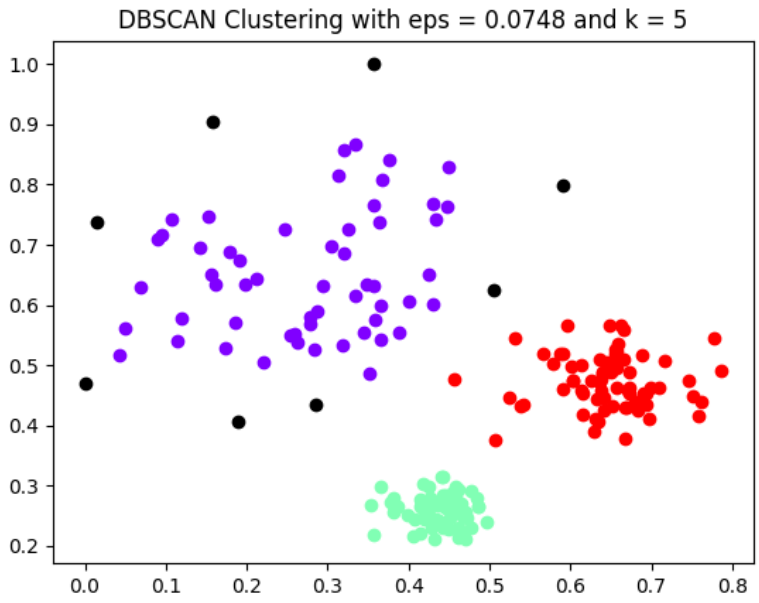

Figure 5: *DBSCAN with* $\epsilon = 0.0748$ *and* $k = 5$

Figure 5 shows the result of the DBSCAN algorithm with $\epsilon = 0.0748$ and $k = 5$. The algorithm found 3 clusters and multiple noise points. The clusters are clearly visible in the figure. This also shows outliers which are not clearly visible in the figure.

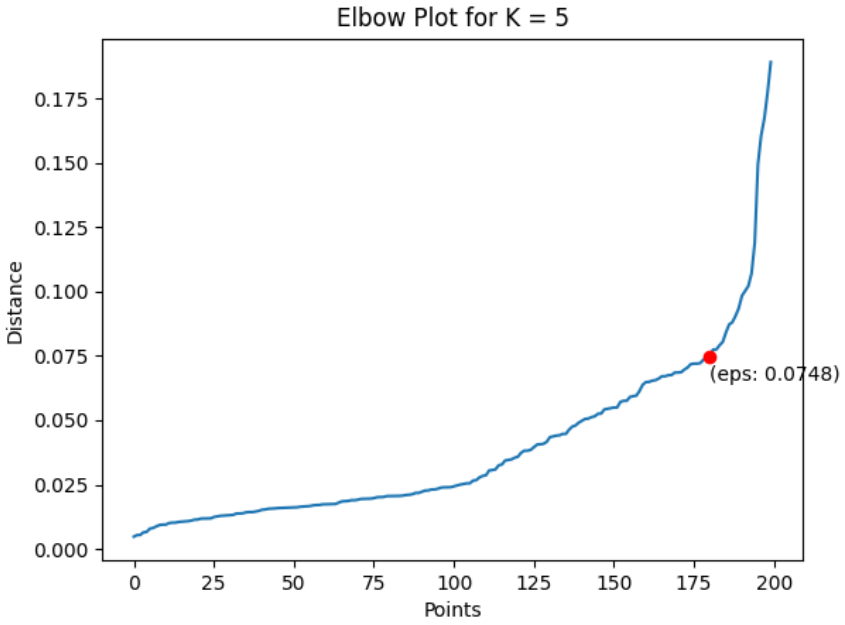

Figure 6: *Elbow plot with $k = 5$*

Figure 6 shows the elbow plot with $k = 5$. The red dot is the elbow point determined by the researchers. This should be the place with the greatest change in the slope of the curve.

k	eps	silhouette score
3	0.0493	0.37633100578468304
4	0.071	0.6507003142496615
5	0.0748	0.6509892268469649

Figure 7: *Silhouette scores*

Figure 7 shows the silhouette scores for the different k values and the *eps* values determined by their elbow plots.

4. Discussion

The silhouette scores show that the best value for $k = 5$. The diference between the scores for $k = 4$ and $k = 5$ is very small. The fact that the researchers determined the elbow points themselves is a disadvantage and could have influenced the results. So realistically there is not a definitive best value for k when comparing 4 and 5. Furthermore, it advised to use 4 as the value for k, which could be a reason to choose $k = 4$. To conclude, the best value for k is either 4 or 5, both values are equally as fair to use.

Bonus

Precision	0.9753446447507953
Recall	1.0
F_measure	0.9875184538988055

Table 1: *Evaluation of the DBSCAN algorithm*

Table 1 shows the evaluation of the DBSCAN algorithm. The precision is 0.9753446447507953, the recall is 1.0 and the F-measure is 0.9875184538988055. The precision is calculated by dividing the number of true positives by the number of true positives and false positives. The recall is calculated by dividing the number of true positives by the number of true positives and false negatives. The F-measure is calculated by dividing the precision and recall by the sum of the precision and recall. The F-measure is a harmonic mean of the precision and recall. The F-measure is a good measure to use when the precision and recall are not equally important.

Work distribution

This week the work has been divided relatively equally between the group members. Ramon has written the majority of the code and Stijn has finished it off after together examining the code and functionality of it. For the report, Stijn has written the methods used and the majority of the results. Ramon Has written the discussion. All other parts have been written together and the final report is scanned by both Ramon and Stijn and they have discussed and implemented the possible changes.