Introduction to Machine Learning
Lab Session 3

Group 39
Laura Lall (s6221858) & Rachel Yang (s6213154)

March 5, 2025

Introduction

This report presents the implementation and analysis of Density-based spatial clustering of applications with noise (DBSCAN) algorithm on a dataset containing 200 two-dimensional feature vectors. The objective is to cluster data points and detect outliers using the Euclidean distance optimizing the parameter eps based on the k-nearest neighbor (k-NN) method for different values of MinPts.

Methods

DBSCAN groups closely packed points into clusters, while identifying sparse points as noise. It relies on these two parameters:

eps - the maximum distance between two points, at which they are considered neighbours

MinPts - the minimum number of points required to form a dense region.

The algorithm starts by marking all points as unvisited. For each unvisited point a regionQuery is performed to find all points that fall within eps. If the number of neighbours for a point is less than MinPts, the point is marked as noise, otherwise a new cluster is created and expanded by using the expandCluster function. The algorithm continues until all points are visited.

Experimental results

Qualitative:

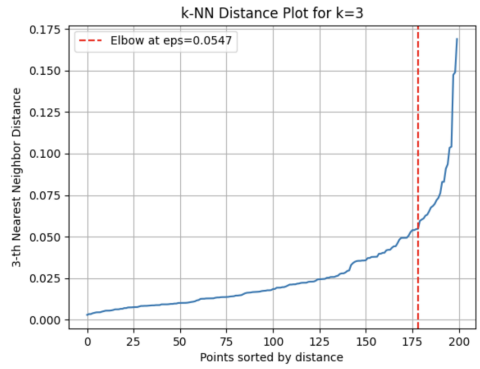

Figure 1: *KNN with MinPts=3*

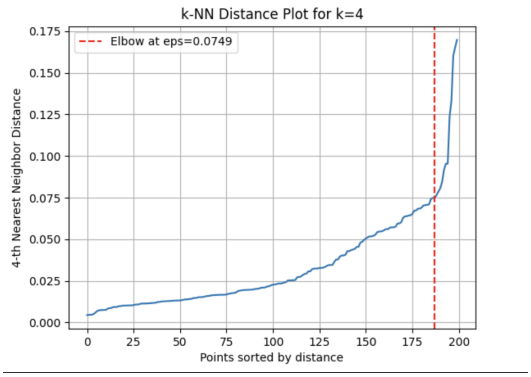

Figure 2: *KNN with MinPts=4*

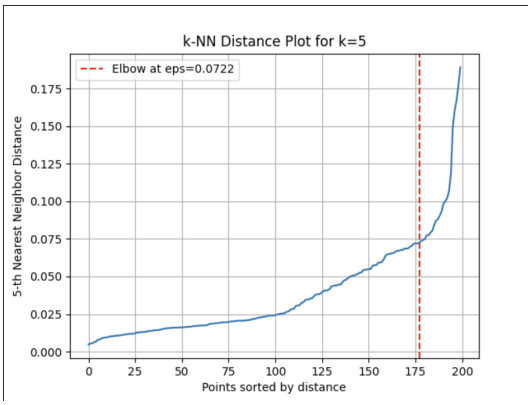

Figure 3: *KNN with MinPts=5*

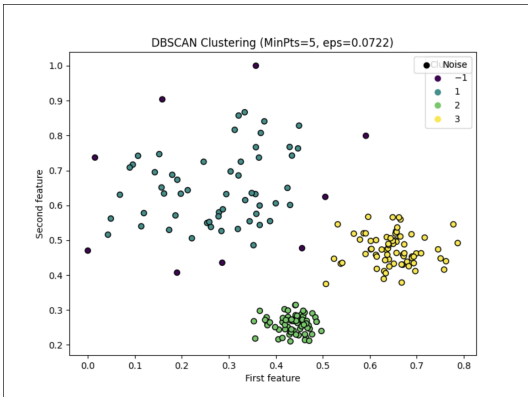

Figure 4: *DBSCAN with MinPts=5, eps=0.0722*

Quantitative:

eps	MinPts	Silhouette score
0.0547	3	0.4308
0.0749	4	0.6045
0.0722	5	0.6057

Table 1: *Silhouette scores for different parameter settings*

Discussion

Qualitative: For MinPts = 3, the algorithm produced well-defined clusters with a clear separation between dense regions and outliers. The elbow method helped identify an optimal eps value, which resulted in meaningful clusters. For MinPts = 4, the number of clusters decreased slightly, and some points that were previously part of smaller clusters were classified as noise. This indicates that increasing MinPts makes the algorithm more conservative in forming clusters. For MinPts=5, the algorithm became even more restrictive, resulting in fewer clusters and more points being classified as noise. This setting may be useful in scenarios where the goal is to identify only the most dense regions.

Quantitative: The silhouette scores were highest for MinPts=5, indicating that this setting produced the most well-separated and cohesive clusters. The optimal eps value, determined using the elbow method, contributed to this high score. For MinPts=4, the silhouette score was slightly lower, but the clusters were still well-defined. This suggests that while decreasing MinPts can increase the number of clusters, it might not significantly degrade cluster quality. For MinPts=3, the silhouette score decreased further, indicating that the clusters were less distinct. This is likely because the lower MinPts value caused more points to be classified into a cluster, reducing the overall cluster quality.

Conclusion: Based on both qualitative and quantitative results, the optimal parameter settings for this dataset are eps = 0.0722 and MinPts = 5, as this configuration produced the highest silhouette score, indicating the best cluster cohesion and separation. However, MinPts = 4 is a viable alternative, as it maintains strong clustering performance while allowing more points to be included in the clusters. In contrast, MinPts = 3 results in lower cluster quality due to excessive inclusiveness, leading to less distinct clusters. Thus, MinPts = 5 and eps = 0.0722 are recommended for achieving well-separated and compact clusters, while MinPts = 4 may be considered when a slightly more inclusive clustering approach is needed.