**Data Mining and Discovery**

**Assignment Report**

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density-based unsupervised clustering technique. Clusters in DBSCAN are built from dense regions and separated by no or low-density regions. In contrast to k-means clustering, which commonly produces spherical-shaped clusters, DBSCAN computes closest neighbor networks and generates arbitrary-shaped clusters in datasets (which may contain noise or outliers). DBSCAN, unlike k-means clustering, does not require an initial specification of the number of clusters. DBSCAN, on the other hand, requires two parameters to build clusters: the radius of neighborhoods for a given data point p (eps) and the minimum number of data points in a specific - neighborhood (minPts).

A border point is a data point that is within p's -neighborhood and has fewer than the minimal number of points (minPts) within its -neighborhood. A noise point is a point that is not a core point or a boundary point (Outlier). If point p is a core point and x is in p's -neighborhood, a point x is directly density accessible from point p. If a point y is directly density accessible to core point x, which is likewise density reachable to core point p, it is density reachable from point p.

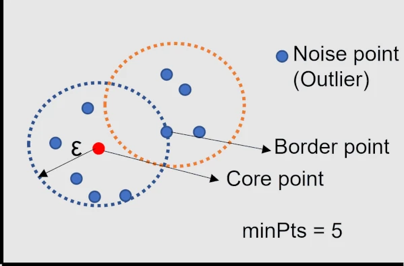


Figure DBSCAN technology

All locations inside a core point's -neighborhood belong to the same clusters, and all places within a cluster are directly density accessible from the core point. The sites that cannot be reached by density from any core point are referred to as noise (outliers). The steps involved in DBSCAN clustering algorithm are as follows:

1. Choosing a point randomly (p).
2. Identify density reachable points from point p with e and minPts parameter.
3. If the core point is P, create a cluster with e and minPts.
4. If the border point is p, visit the next point in a dataset.
5. Continue the algorithm until all points are visited.

Basketball dataset was chosen for the DB clustering algorithm mainly for two main factors. First, numerical data produces very clear data clustering compared to string-based data and second, passing ability of an individual indicates amount of training the team needs. The data was extracted using an app by the data extractors. The data used in this clustering is scoped to a single individual and his passing parameters indicating the acceleration in x direction and the turning degree of the ball.

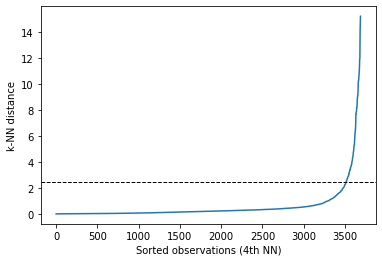


Figure Graph to identify the 'e' parameter

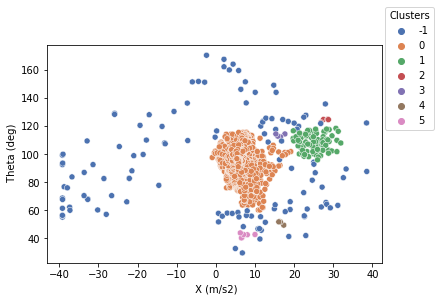
Figure 1 above illustrates the intial step in graph-based clustering which is finding the value of k-NN (eps) to identify the noise and outliers. It is clear that the knee occurs at approximately 2.5 and thus points below 2.5 belong to acluster and above are noises.

Figure Data clustered using DBSCAN

With respect to the values for ‘e’ and ‘minPts’, Figure 2 illustrates how the data is being clustered. The legend shows that there are five clusters of data and -1 indicates noisy points which cannot be assigned to any cluster. The orange cluster (Cluster 0) depicts the individuals most occurring angle and acceleration of his ball pass. This is useful for him as well as the coach to estimate the players performance and is a good way to measure the players progress with time.

**References**

Bedre, R. (2022) *DBSCAN in python (with example dataset)*, *Data science blog*. Available at: https://www.reneshbedre.com/blog/dbscan-python.html (Accessed: January 6, 2023).

Chabot, O. (no date) *10 spatial clustering: Spatial analysis of basketball shots in R*, *10 Spatial Clustering | Spatial Analysis of Basketball Shots in R*. Available at: https://duddhawork.com/ballbook/clusters.html (Accessed: January 6, 2023).

*DBSCAN clustering in ML: Density based clustering* (2022) *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ (Accessed: January 6, 2023).

*Demo of DBSCAN clustering algorithm* (no date) *scikit*. Available at: https://scikit-learn.org/stable/auto\_examples/cluster/plot\_dbscan.html (Accessed: January 6, 2023).

Zhang, Y. (2022) *DBSCAN clustering algorithm based on big data is applied in Network Information Security Detection*, *Security and Communication Networks*. Hindawi. Available at: https://www.hindawi.com/journals/scn/2022/9951609/ (Accessed: January 6, 2023).