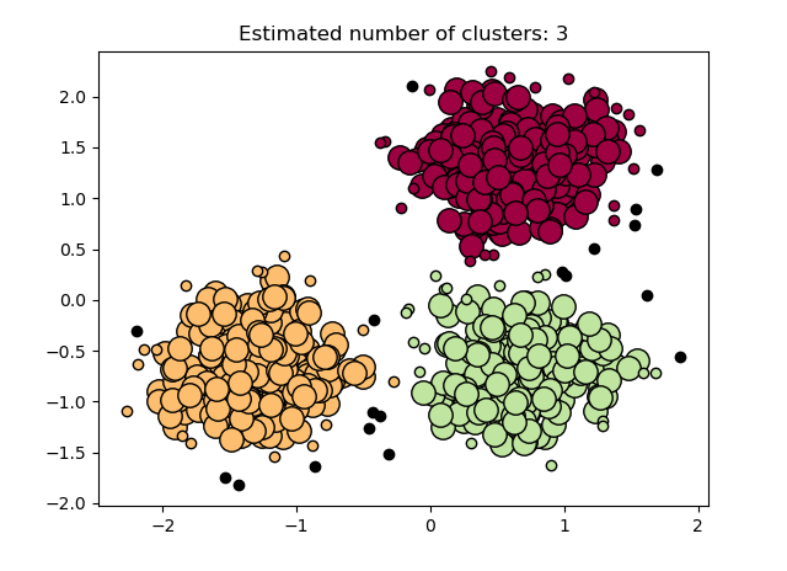
Machine Learning Clustering

A. BARKATHULLA

Feb-2025

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- DBSCAN is a density-based clustering algorithm that is used to identify clusters in a dataset.
- DBSCAN does not require the number of clusters to be specified beforehand.
- Instead, it uses a density-based approach to identify clusters of similar data points.



How Does DBSCAN Work?

- The basic idea behind DBSCAN is to identify clusters of data points that are densely packed together.
- It starts by selecting a random data point and then finding all data points within a specified radius (referred to as the “eps” parameter).
- If there are enough data points within this radius (referred to as the “minPts” parameter), a cluster is created and the algorithm continues to find all data points within the same radius.
- The algorithm repeats this process for each data point in the cluster until no more data points can be added.

Advantages of DBSCAN

1. Can handle datasets with varying cluster sizes and shapes.
2. Does not require the number of clusters to be specified beforehand, making it a more flexible clustering algorithm.
3. Can identify clusters of arbitrary shapes, which makes it well-suited for datasets with complex structures.
4. Can be used in a wide range of applications, including image segmentation, customer segmentation, and gene expression analysis.

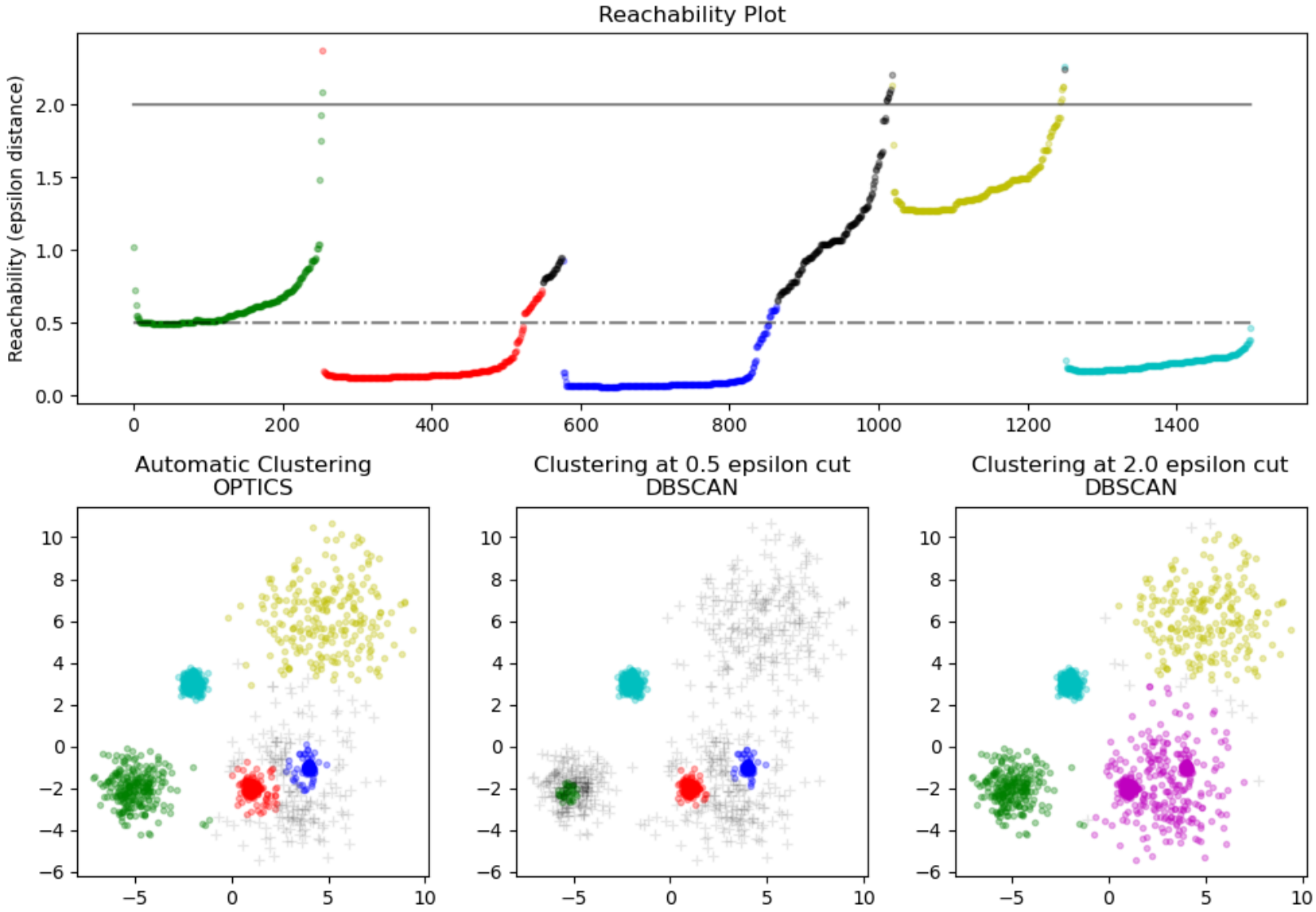
Disadvantages of DBSCAN

- Can be sensitive to the choice of the “eps” and “minPts” parameters, making it important to choose these values carefully.
- Can be computationally expensive, especially for large datasets.
- Can be affected by the presence of noise and outliers in the dataset.
- DBSCAN cannot cluster data sets well with large differences in densities, since the “eps” and “minPts” parameters are the same for all clusters.

OPTICS (Ordering Points To Identify Cluster Structure)

- OPTICS is a density-based clustering algorithm, similar to DBSCAN, but it can extract clusters of varying densities and shapes.
- The main idea behind OPTICS is to extract the clustering structure of a dataset by identifying the density-connected points.
- The algorithm builds a density-based representation of the data by creating an ordered list of points called the reachability plot.
- Each point in the list is associated with a reachability distance, which is a measure of how easy it is to reach that point from other points in the dataset.
- Points with similar reachability distances are likely to be in the same cluster.

OPTICS (Ordering Points To Identify Cluster Structure)



How Does OPTICS Work?

1. **Memory Cost :** The OPTICS clustering technique requires more memory as it maintains a priority queue (Min Heap) is used to find Reachability Distance. Where as DBSCAN requires less memory space.
2. **Handling varying densities:** OPTICS can identify clusters of different sizes and shapes more effectively than DBSCAN in datasets with varying densities.
3. **Noise handling:** OPTICS may be less effective when compared to DBSCAN at identifying small clusters that are surrounded by noise points, as these clusters may be merged with the noise points in the reachability distance plot.
4. **Runtime complexity:** The runtime complexity of OPTICS is generally higher than that of DBSCAN
5. **Fewer Parameters:** OPTICS has fewer parameters when compared to DBSCAN

Advantages of OPTICS

- OPTICS clustering doesn't require a predefined number of clusters in advance
- Clusters can be of any shape, including non-spherical ones

Disadvantages of OPTICS

- It fails if there are no density drops between clusters
- It is also sensitive to parameters that define density (radius and the minimum number of points) and proper parameter settings require domain knowledge.

BIRCH -Balanced Iterative Reducing and Clustering using Hierarchies -

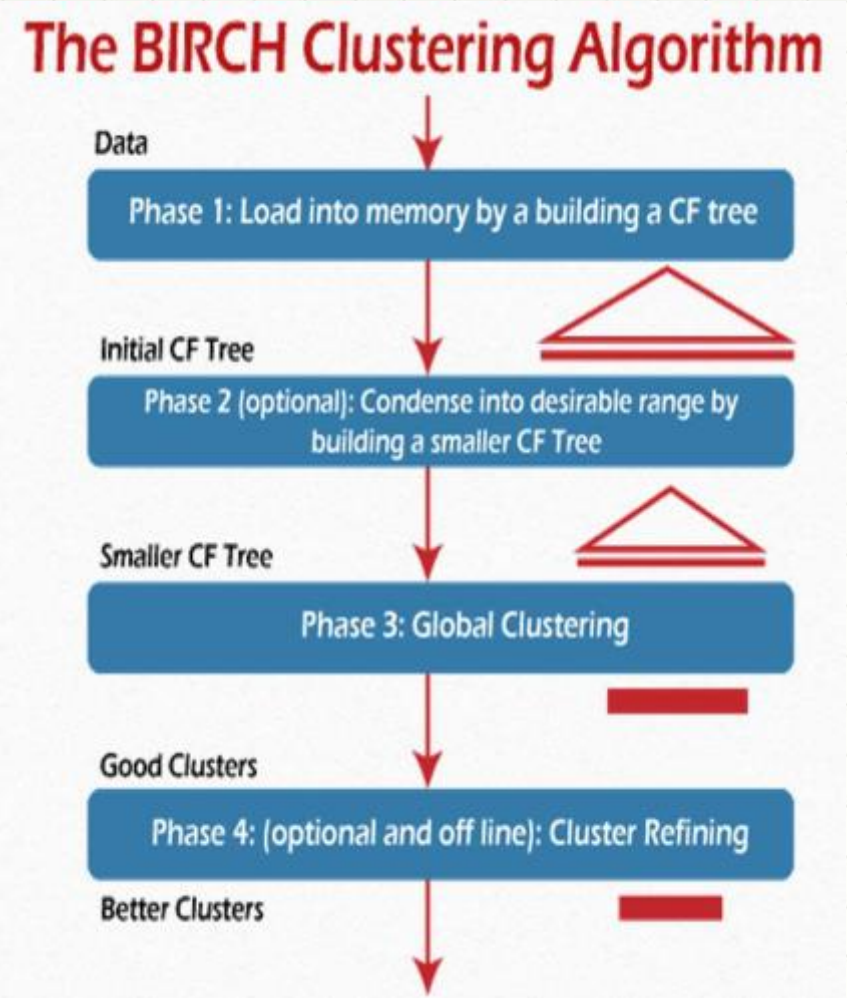
- BIRCH is a hierarchical clustering algorithm that is designed to handle large datasets efficiently.
- The algorithm builds a treelike structure of clusters by recursively partitioning the data into sub clusters until a stopping criterion is met.
- BIRCH uses two main data structures to represent the clusters: Clustering Feature (CF) and Sub-Cluster Feature (SCF).
- CF is used to summarize the statistical properties of a set of data points, while SCF is used to represent the structure of sub clusters.

How Does BIRCH works?

1. **Initialization** – BIRCH constructs an empty tree structure and sets the maximum number of CFs that can be stored in a node.
2. **Clustering** – BIRCH reads the data points one by one and adds them to the tree structure.
   - If a CF is already present in a node, BIRCH updates the CF with the new data point.
   - If there is no CF in the node, BIRCH creates a new CF for the data point. BIRCH then checks if the number of CFs in the node exceeds the maximum threshold.
   - If the threshold is exceeded, BIRCH creates a new sub cluster by recursively partitioning the CFs in the node.
3. **Refinement** – BIRCH refines the tree structure by merging the sub clusters that are similar based on a distance metric.

How Does BIRCH works?

- .



Advantages of BIRCH

- **Scalability** – BIRCH is designed to handle large datasets efficiently by using a treelike structure to represent the clusters.
- **Memory efficiency** – BIRCH uses CF and SCF data structures to summarize the statistical properties of the data points, which reduces the memory required to store the clusters.
- **Fast clustering** – BIRCH can cluster the data points quickly because it uses an incremental clustering approach.

Disadvantages of BIRCH

- **Sensitivity to parameter settings** – The performance of BIRCH clustering can be sensitive to the choice of parameters, such as the maximum number of CFs that can be stored in a node and the threshold value used to create subclusters.
- **Limited ability to handle non-spherical clusters** – BIRCH assumes that the clusters are spherical, which means it may not perform well on datasets with nonspherical clusters.
- **Limited flexibility in the choice of distance metric** – BIRCH uses the Euclidean distance metric by default, which may not be appropriate for all datasets.