Machine Learning Clustering

A. BARKATHULLA

Feb-2025

Unsupervised Clustering

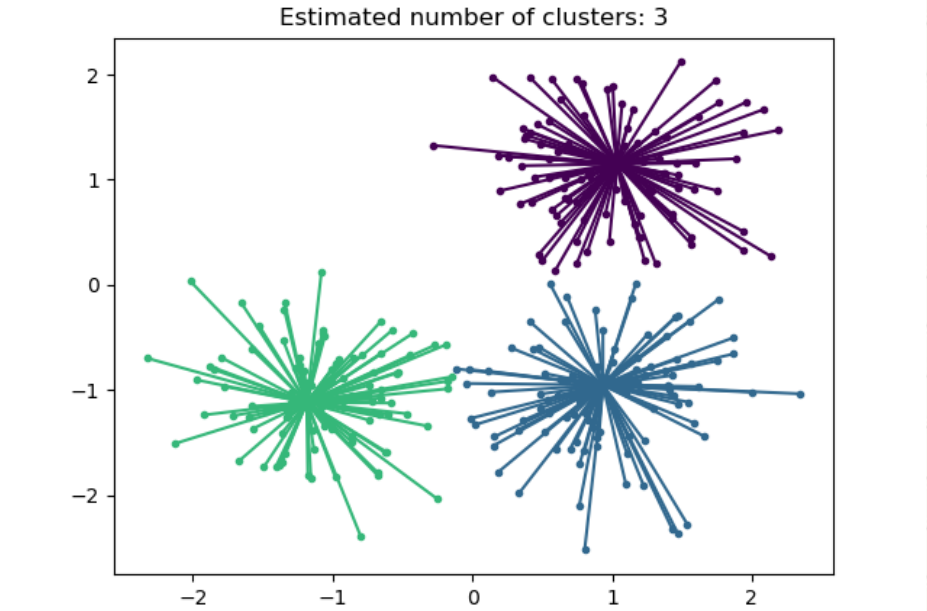
- Unsupervised Clustering is a type of machine learning technique .
- Where a computer program is used to automatically group similar objects or data points together without any prior knowledge of what theses group should be got .

Unsupervised Clustering

- The goal of unsupervised clustering is to identifying natural grouping within the data
- Which can be used for further analysis or decision making some common applications of unsupervised clustering.
- Include market segmentation image and document classification and anomaly detection.

AFFINITY PROPAGATION ALGORITHM

- Affinity Propagation is a clustering algorithm used to cluster data points into multiple groups based on their similarity.



How Does Affinity Propagation Algorithm work?

- Affinity Propagation does not require the number of clusters to be specified beforehand.
- Instead, it iteratively adjusts the “responsibilities” and “availabilities” between data points to determine the number of clusters and the assignment of data points to those clusters.

How Does Affinity Propagation Algorithm work?

- The key idea behind Affinity Propagation is that each data point can both act as an exemplar (representative) of its own cluster, and can also have a preference for being the exemplar of other data points.
- The algorithm seeks to find the exemplars that result in the highest total preference among all data points.

How Does Affinity Propagation Algorithm work?

- The algorithm can be applied to a wide range of applications, including image segmentation, customer segmentation, and gene expression analysis.
- However, it can be computationally expensive, especially for large datasets, and may not always produce the best results compared to other clustering algorithms.

Advantages of Affinity Propagation

1. Does not require the number of clusters to be specified beforehand, making it a more flexible clustering algorithm.
2. Can produce high-quality clusters even when the data points have different densities or sizes.
3. Can be used to cluster data with complex relationships and non-linear structures.
4. Can be used in a wide range of applications, including image segmentation, customer segmentation, and gene expression analysis.

Disadvantages of Affinity Propagation

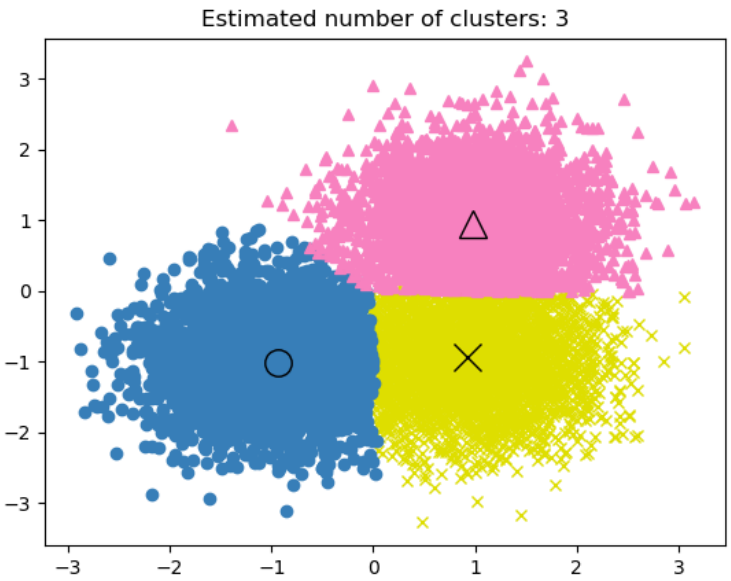
1. Can be computationally expensive, especially for large datasets, making it unsuitable for large-scale clustering problems.
2. May not always produce the best results compared to other clustering algorithms, such as K-Means or Gaussian Mixture Models.
3. Can be sensitive to the choice of similarity metric used to measure the similarities between data points.
4. Can produce multiple exemplars for a single cluster, making it difficult to interpret the results of the clustering process.

Conclusion of Affinity Propagation

1. Affinity Propagation is a powerful clustering algorithm that can be used to cluster data points into multiple groups based on their similarity.
2. Despite its strengths, it can be computationally expensive and may not always produce the best results compared to other clustering algorithms.
3. However, it can still be a useful tool in many applications, especially when the number of clusters is not known beforehand or when the data points have complex relationships.

MEAN-SHIFT CLUSTERING ALGORITHM

- Mean-shift clustering is a powerful and flexible non-parametric clustering algorithm.
- It has found applications in various domains such as computer vision, image processing, and bioinformatics.



HOW DOES MEAN-SHIFT CLUSTERING WORK?

- Mean-shift clustering operates by shifting the mean of a set of points in the feature space until it converges to a dense region, called a mode.
- Each mode found by the algorithm is considered to be a cluster.
- The algorithm is based on the idea of kernel density estimation, which is a way of estimating the probability density function of a data set.

HOW DOES MEAN-SHIFT CLUSTERING WORK?

- The first step of the mean-shift algorithm is to define a kernel function, which is used to compute the probability density of the data points.
- A common choice for the kernel function is a Gaussian function.

HOW DOES MEAN-SHIFT CLUSTERING WORK?

- The next step is to initialize the mean of the data points and start the iterative process of shifting the mean towards the dense regions in the data set.

HOW DOES MEAN-SHIFT CLUSTERING WORK?

- Each iteration, the algorithm computes the probability density of the data points given the current mean.
- Then, the mean is shifted to the weighted average of the data points, where the weights are determined by the probability density.
- This process is repeated until the mean converges or a maximum number of iterations is reached.
- The final mean is assigned to a cluster, and the data points that are closest to the mean are assigned to that cluster as well.

HOW DOES MEAN-SHIFT CLUSTERING WORK?

We can understand the working of Mean-Shift clustering algorithm with the help of following steps

- Step 1 − First, start with the data points assigned to a cluster of their own.
- Step 2 − Next, this algorithm will compute the centroids.
- Step 3 − In this step, location of new centroids will be updated.
- Step 4 − Now, the process will be iterated and moved to the higher density region.
- Step 5 − At last, it will be stopped once the centroids reach at position from where it cannot move further.

Advantages mean-shift clustering

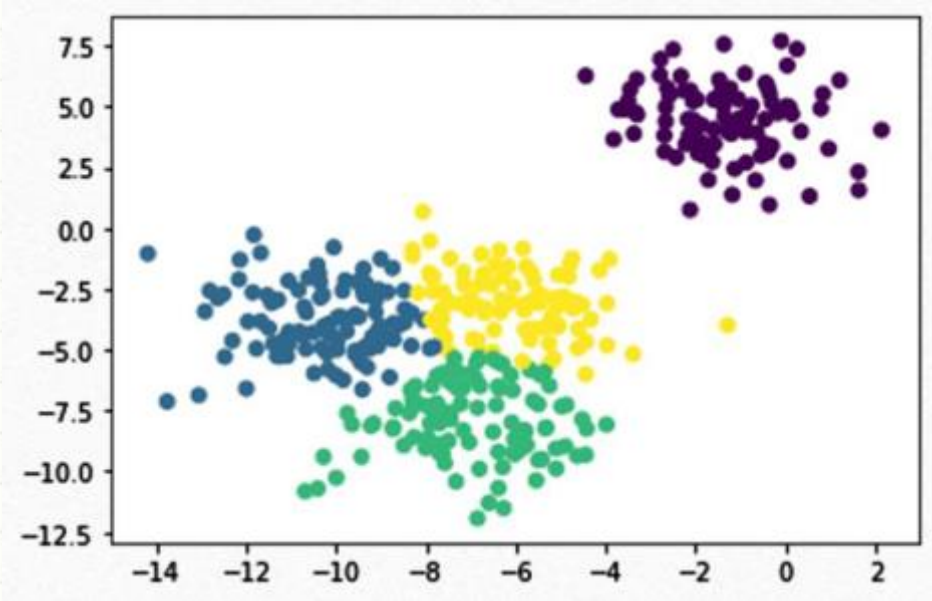
- Mean-shift algorithm is that it does not require the number of clusters to be specified beforehand, unlike other algorithms such as K-Means.
- This makes it an attractive option when the number of clusters is unknown or difficult to determine.
- Another advantage is that the mean-shift algorithm can handle non-linearly separated data, which is not the case for algorithms such as K-Means.

Disadvantages mean-shift clustering

- It can be computationally expensive compared to other clustering algorithms. This is because the algorithm requires multiple iterations to converge to the modes.
- Additionally, the algorithm is sensitive to the choice of the bandwidth parameter, which controls the size of the neighborhood considered for mean shifting.
- A smaller bandwidth may result in over-segmentation, while a larger bandwidth may result in under-segmentation.

Spectral Clustering

- Spectral Clustering algorithm is based on the eigenvectors and eigenvalues of the graph Laplacian matrix and works by transforming the data into a lower-dimensional space before clustering it.
- Spectral Clustering is a powerful method for clustering non-linearly separable data, and it is often used.



How Does Spectral Clustering Work?

- The first step is to create a similarity matrix based on the data. This matrix measures the similarity between pairs of data points, and it is used to construct the graph Laplacian matrix.
- The second step is to find the eigenvectors of the graph Laplacian matrix and use them to transform the data into a lower-dimensional space.
- The transformed data is then clustered using a traditional clustering algorithm like K-Means.

Advantages of Spectral Clustering

- **Flexibility:** Spectral clustering is flexible and can handle different types of data distributions and shapes, making it a good choice for a wide range of problems.
- **No assumption on cluster shape:** Unlike other clustering algorithms, such as K-Means, Spectral Clustering does not make any assumptions about the shape of clusters. This makes it a good choice for problems where the clusters have irregular shapes.
- **Better handling of noisy data:** Spectral Clustering is less sensitive to noise and can handle noisy data better than some other clustering algorithms.

Advantages of Spectral Clustering

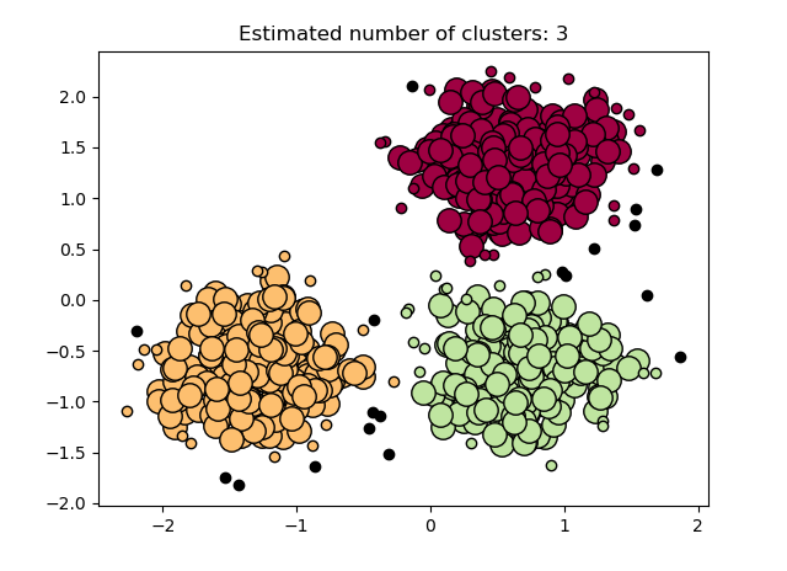
- **Flexibility:** Spectral clustering is flexible and can handle different types of data distributions and shapes, making it a good choice for a wide range of problems.
- **No assumption on cluster shape:** Unlike other clustering algorithms, such as K-Means, Spectral Clustering does not make any assumptions about the shape of clusters. This makes it a good choice for problems where the clusters have irregular shapes.
- **Better handling of noisy data:** Spectral Clustering is less sensitive to noise and can handle noisy data better than some other clustering algorithms.

Disadvantages of Spectral Clustering

- **Scalability:** Spectral Clustering can be computationally expensive, especially for large datasets, as it requires constructing a similarity matrix and solving an eigenvalue problem.
- **Choice of similarity matrix:** The choice of similarity matrix can have a significant impact on the clustering results, and the appropriate choice may not always be clear.
- **Sensitivity to parameter choice:** Spectral Clustering is sensitive to the choice of parameters, such as the choice of kernel or the number of clusters, and poor parameter choices can result in suboptimal clustering results.

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.

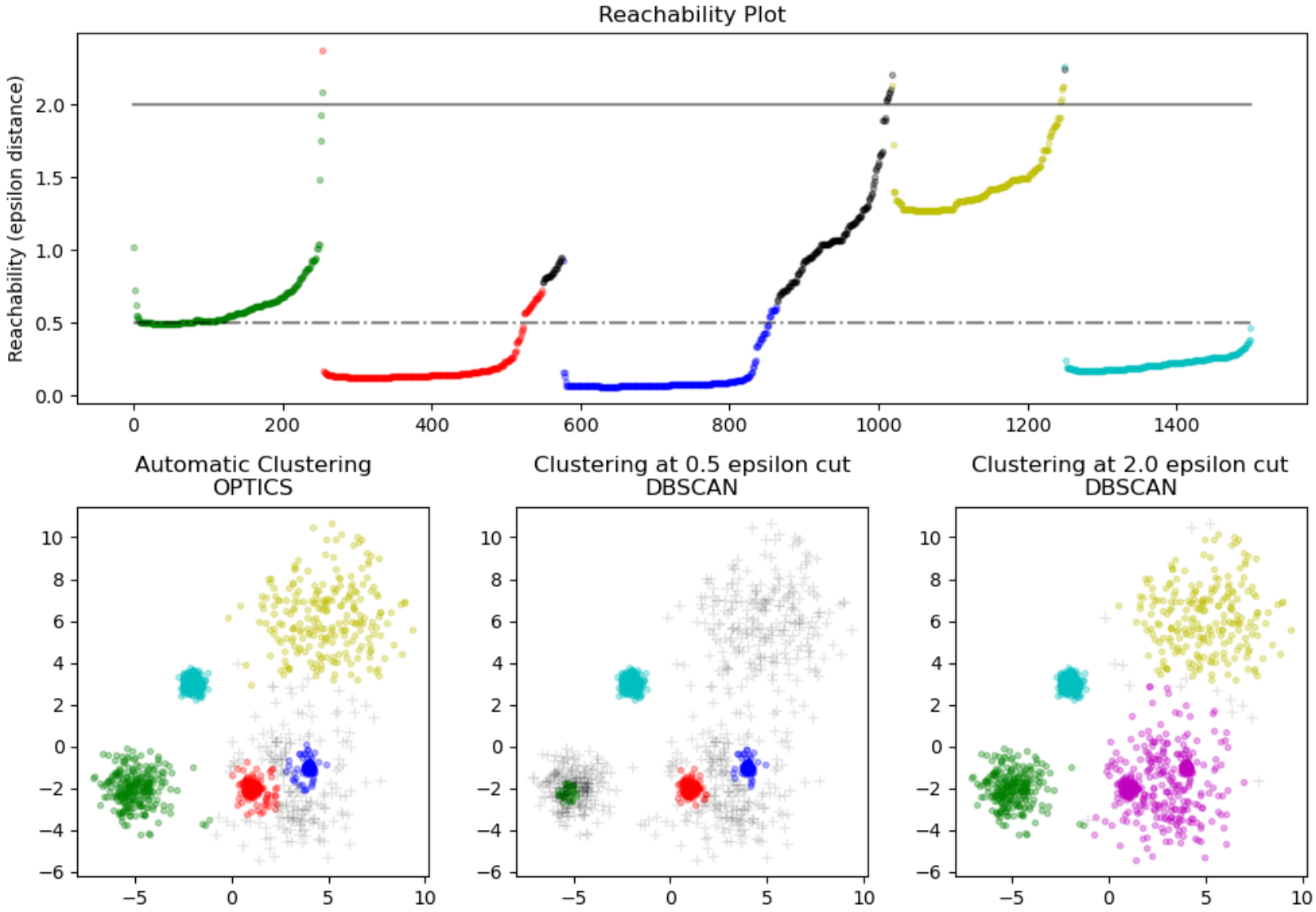
Disadvantages of DBSCAN

- 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 -

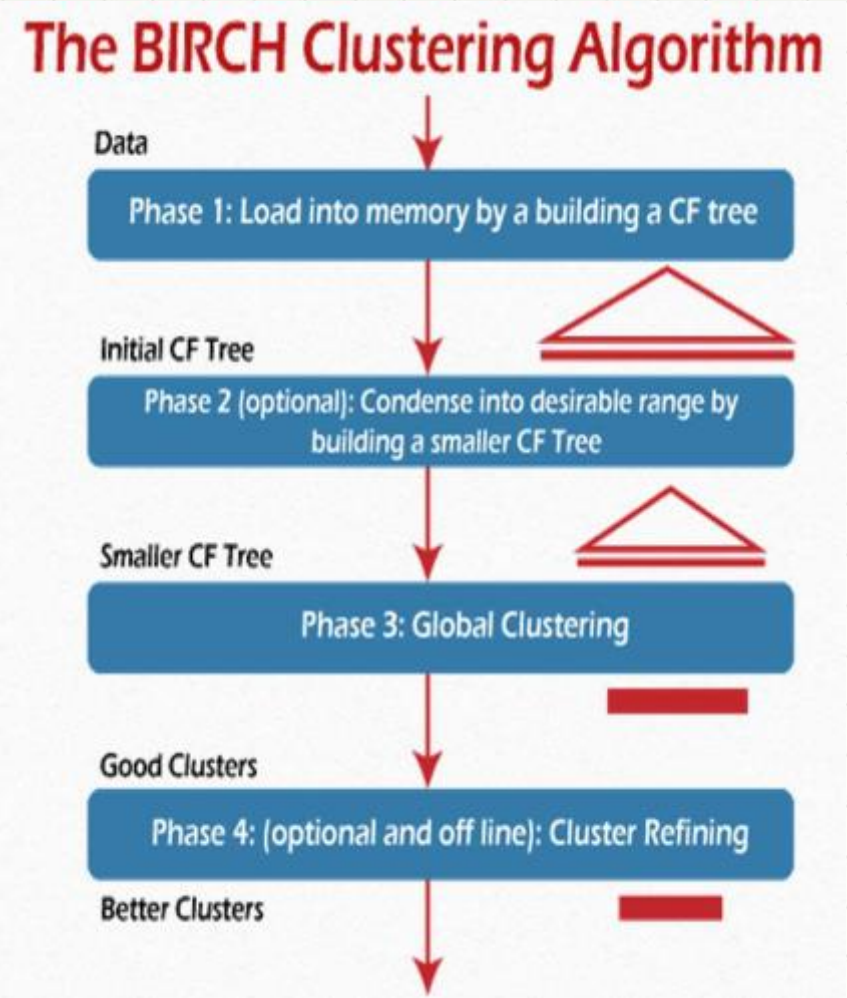
- BIRCH is a hierarchical clustering algorithm that is designed to handle large datasets efficiently.
- The algorithm builds a treelike structure of clusters by recursively partitioning the data into sub clusters until a stopping criterion is met.
- BIRCH uses two main data structures to represent the 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.