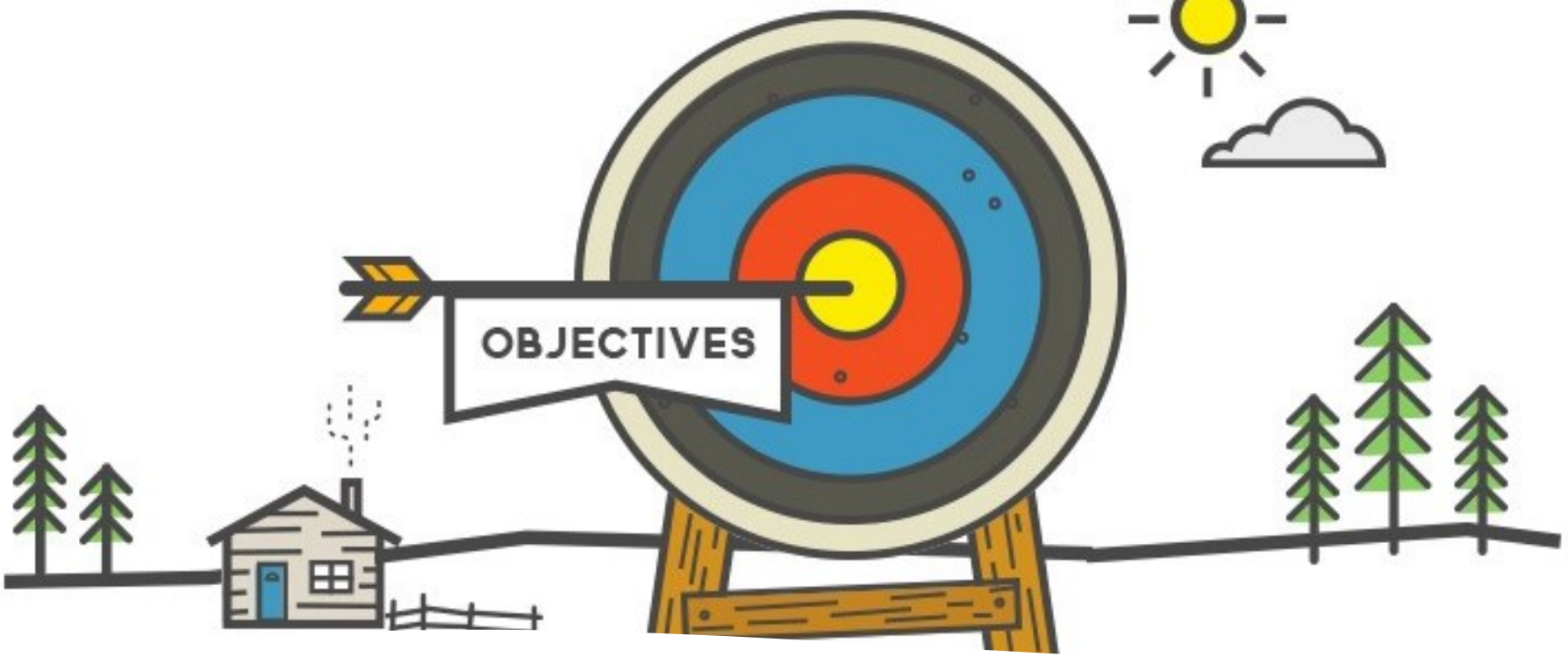
Lisbon School of Economics & Management
Universidade de Lisboa

U LISBOA
UNIVERSIDADE DE LISBOA

Clusters Analysis

Carlos J. Costa

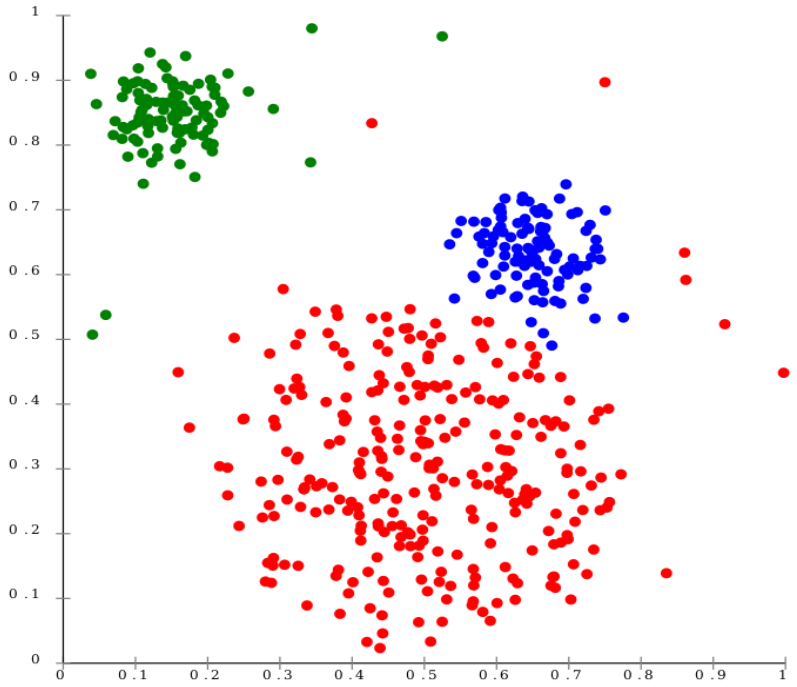


Learning Goals

- Know concept of Cluster analysis
- Distinguish between main algorithms
- Identify main scores used to evaluate the quality of clustering results
- Apply algorithms by using python libraries

Cluster Analysis

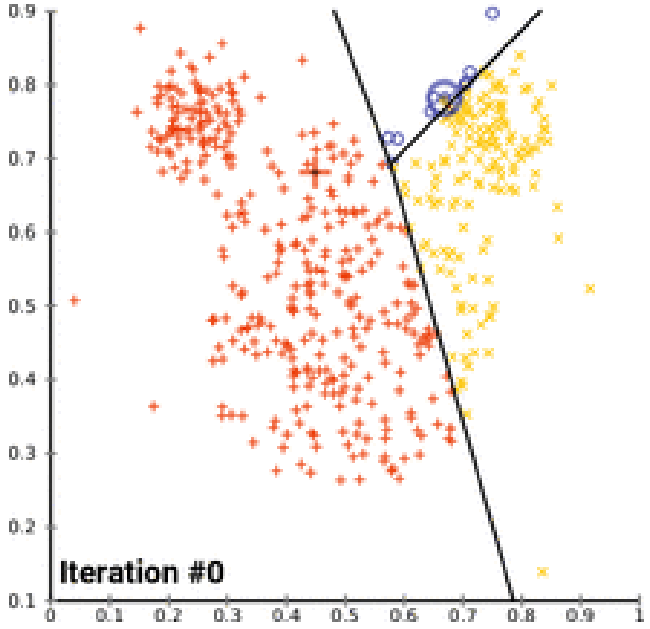
- Cluster analysis is a multivariate method
- aims to classify a sample of subjects (or objects) into several different groups such that similar subjects are placed in the same group
- based on a set of measured variables



Clusters Analysis Types

- Centroid-based Clustering.
- Density-based Clustering.
- Distribution-based Clustering.
- Hierarchical Clustering.

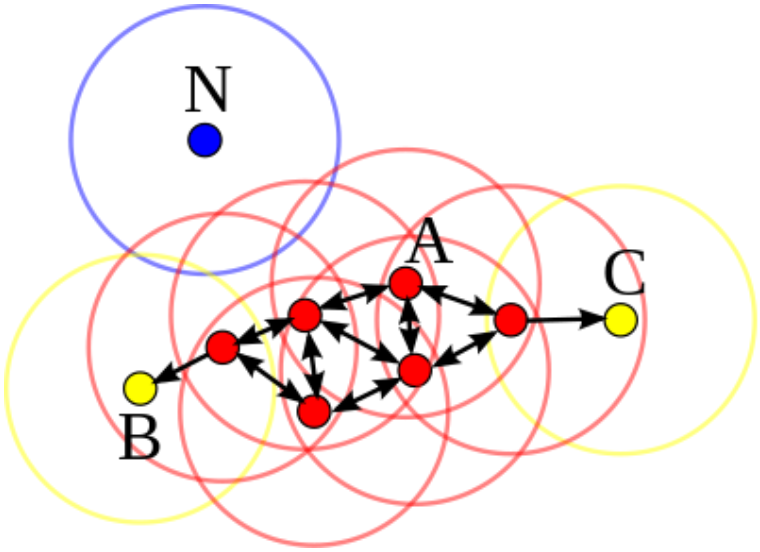
Centroid-based Clustering



- The easiest of all the clustering types in data mining.
- It works on the closeness of the data points to the chosen central value.
- The datasets are divided into a given number of clusters, and a vector of values references every cluster.
- e.g. *k-means clustering, Means Shift Clustering*

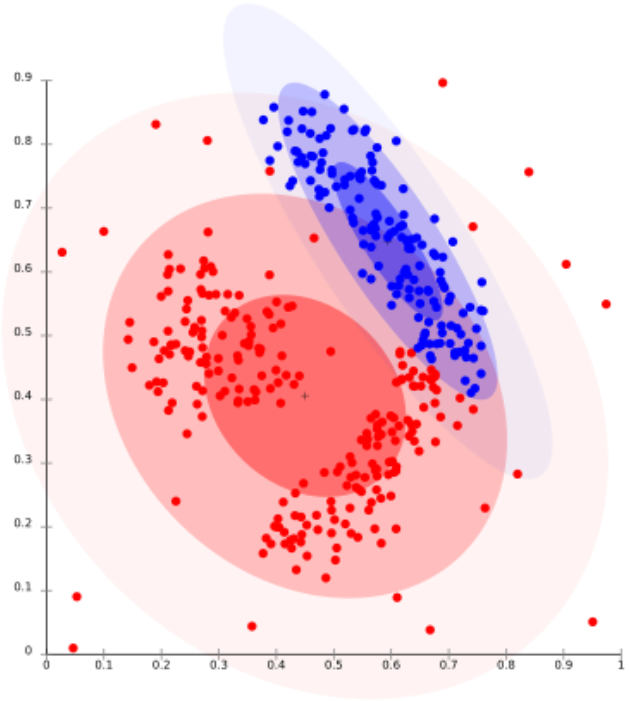
Density-based Clustering

- Considers density ahead of distance.
- Data is clustered by regions of high concentrations of data objects bounded by areas of low concentrations of data objects.
- The clusters formed are grouped as a maximal set of connected data points.
- e.g. DBSCAN and OPTICS



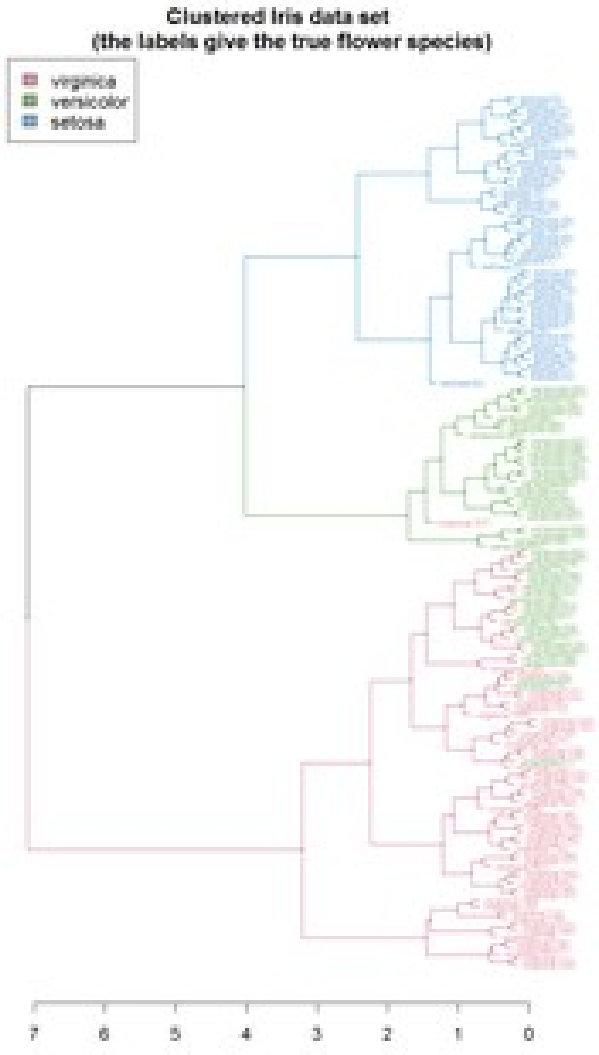
Distribution-based Clustering

- Creates and groups data points based on their likely hood of belonging to the same probability distribution (Gaussian, Binomial, etc.) in the data.
- E.g. expectation-maximization *algorithm*



Hierarchical clustering

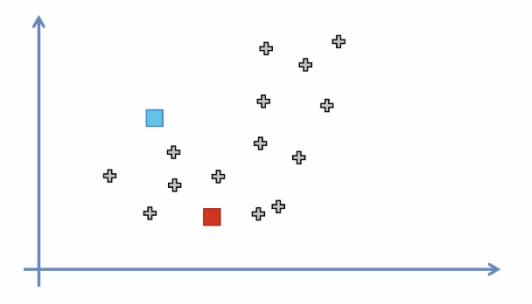
- Based on the principle that every object is connected to its neighbors
- Depending on their proximity distance (degree of relationship).
- e.g. BIRCH (balanced iterative reducing and clustering using hierarchies)



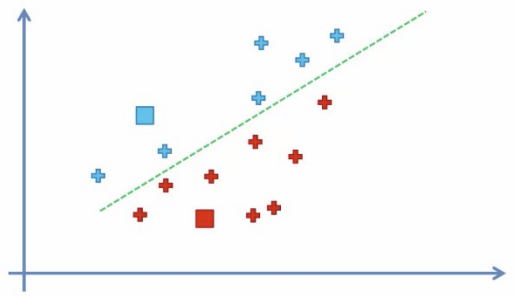
K-means Clustering

Centroid-based Clustering

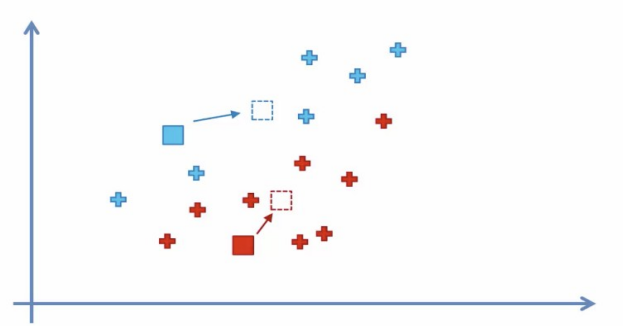
1.Select K (i.e. 2) random points as cluster centres called centroids



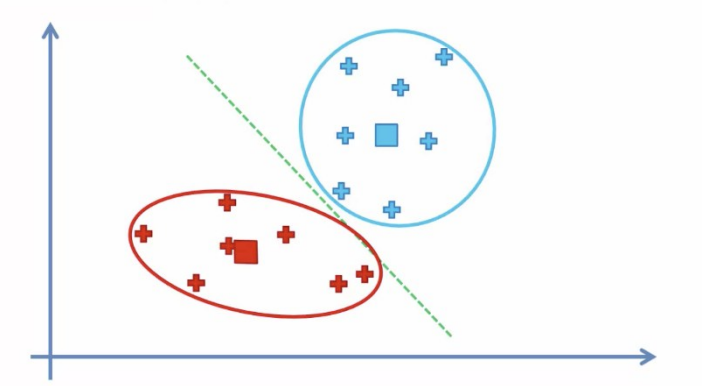
2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid



3. Determine the new cluster centre by computing the average of the assigned points

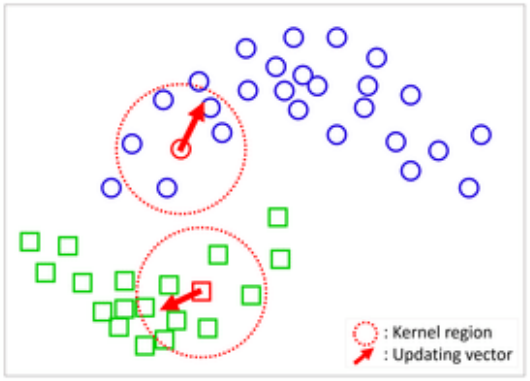


4. Repeat steps 2 and 3 until none of the cluster assignments change



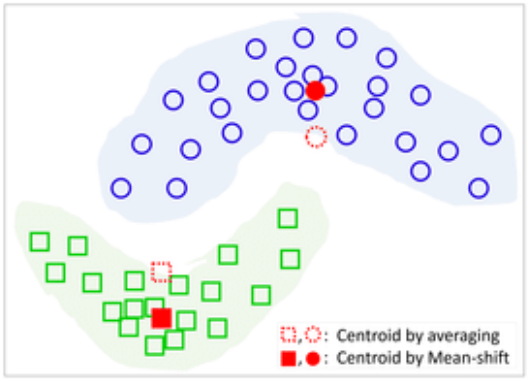
Means Shift Clustering

Centroid-based Clustering



Updates (shifts) all data point toward
high density region until all the points converge

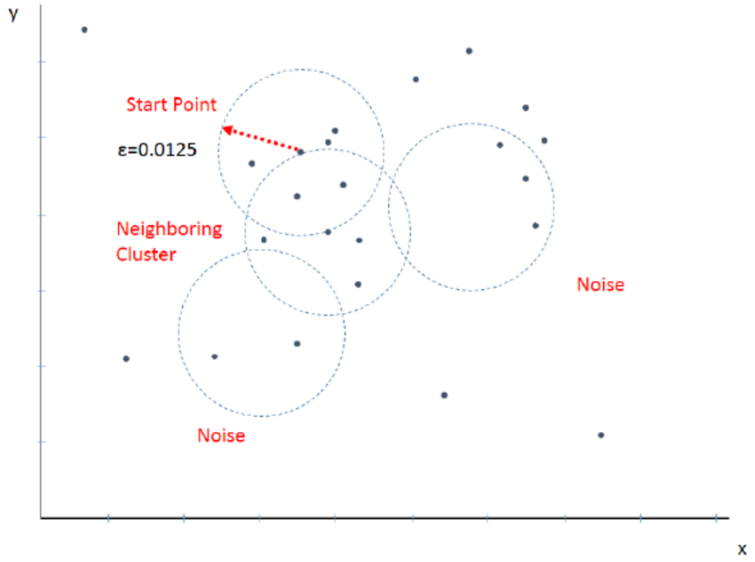
Aggregate the nearby shifted data points
into a cluster whose centroid is their average



Assign the original data into the according clusters,
But keep the centroid calculated with shifted data

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Density-based Clustering
- Groups together data points that are closely packed together and marks points in low-density regions as outliers.
- clusters dense regions of points separated by regions of lower density, without requiring the number of clusters as input.
- Suitable for datasets with varying cluster densities and shapes.

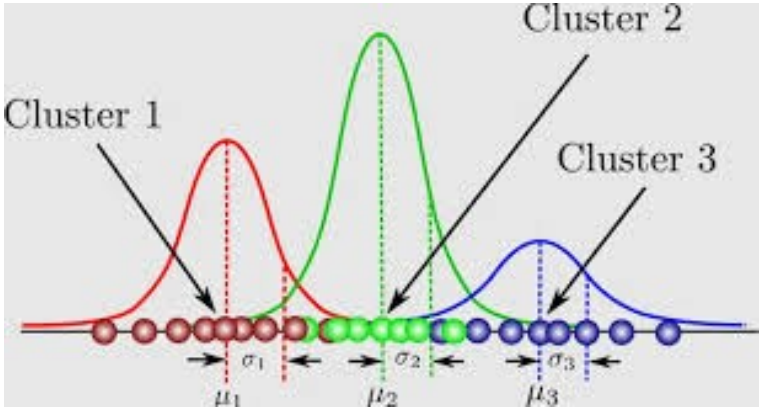


OPTICS

- Ordering Points To Identify the Clustering Structure
- Orders data points based on their density and spatial connectivity, allowing the discovery of clusters of varying density.
- constructs a reachability plot to identify clusters and can handle clusters of arbitrary shapes and sizes.
- Effective for large datasets with noise and varying densities.

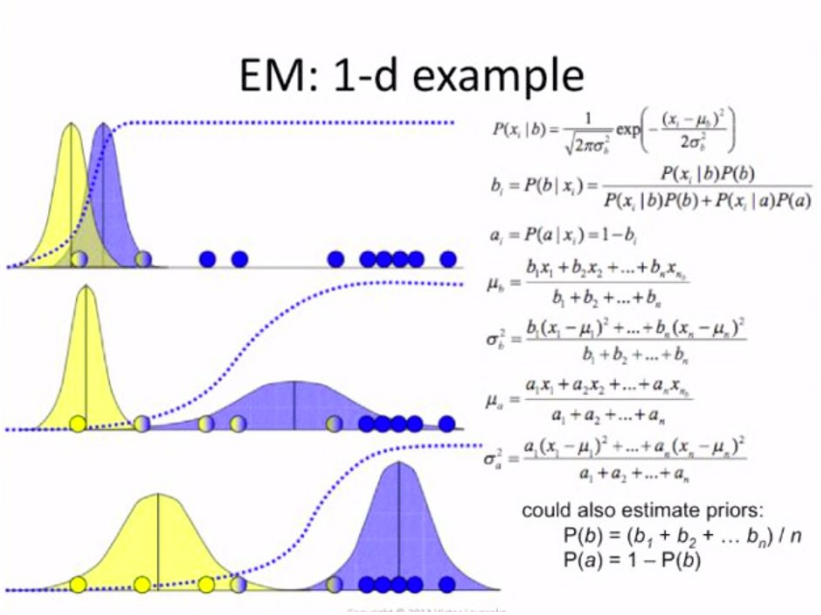
GMM

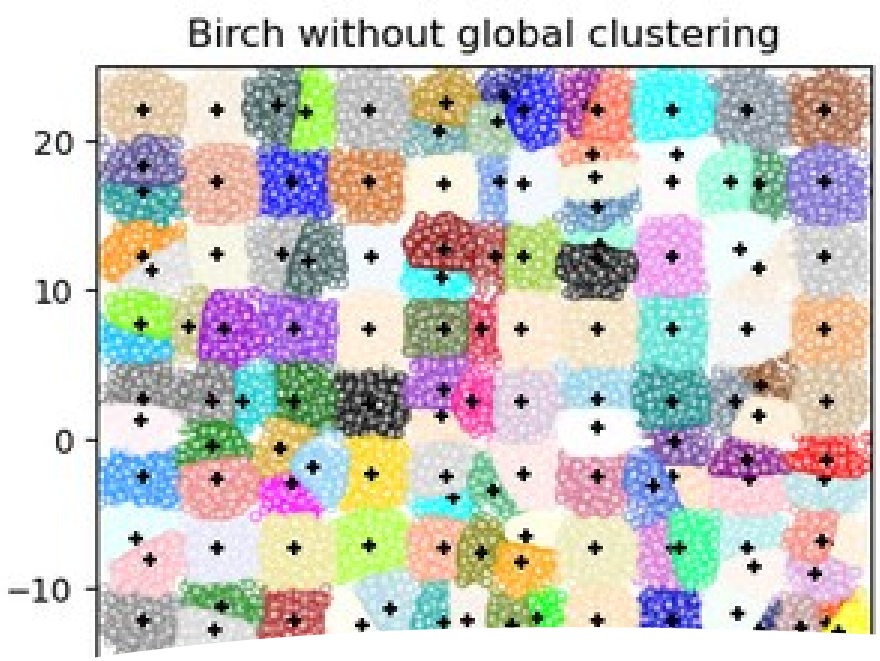
- Gaussian Mixture Models
- Assumes that the data is generated from a mixture of several Gaussian distributions and assigns probabilities to data points belonging to each cluster.
- Models clusters as ellipsoids in the feature space and assigns soft membership to points.
- Suitable for datasets with overlapping clusters.

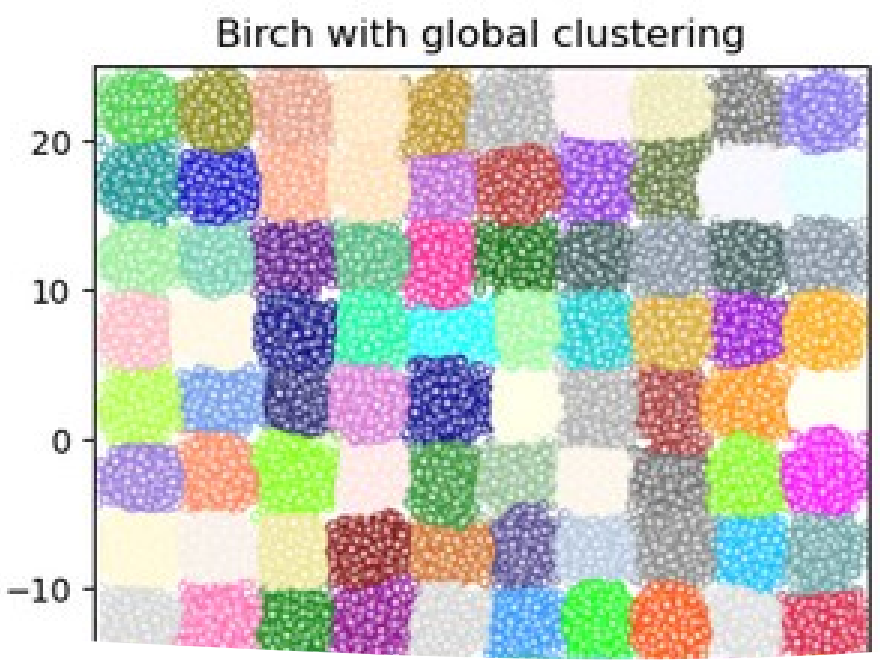


EM Algorithm

- Expectation-Maximization (EM) Algorithm
- A statistical method used to estimate parameters of a mixture model.
- Iteratively estimates the parameters of the model, such as means and covariances, by maximizing the likelihood function of the observed data, often used with GMMs for clustering.
- Effective for datasets with complex data distributions.





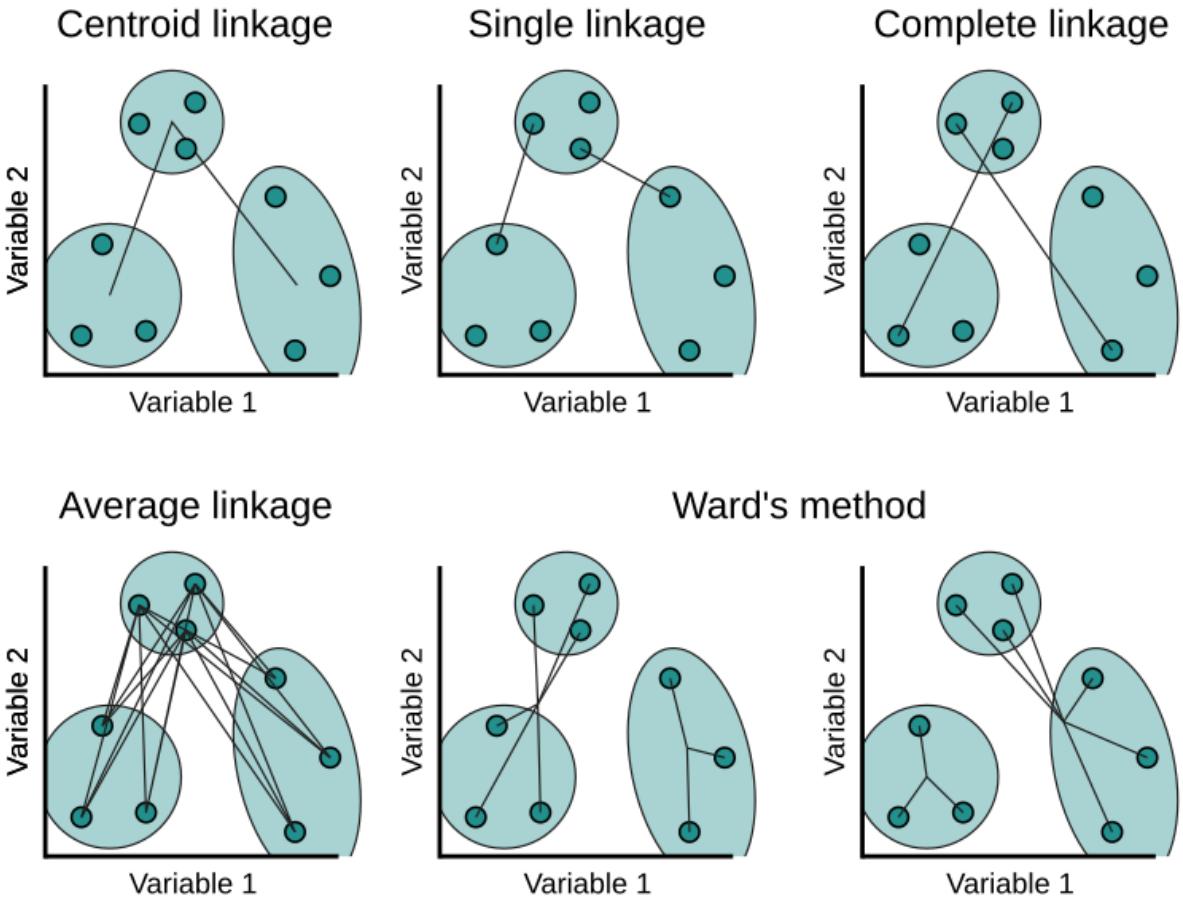


BIRCH

- Balanced Iterative Reducing and Clustering using Hierarchies
- Constructs a hierarchical clustering structure by creating a tree of clusters, known as the Clustering Feature Tree (CF Tree).
- It incrementally processes data points and merges them into clusters while maintaining a compact and efficient representation.
- Ideal for handling large datasets efficiently.

Agglomerative Clustering

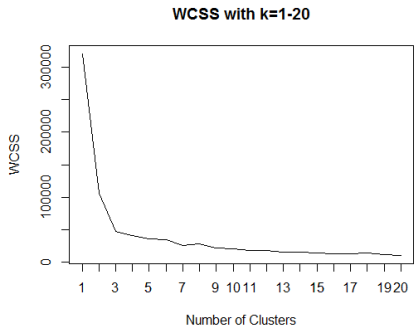
- Hierarchical Clustering Algorithm
- Recursively merges pair of clusters of sample data
- Uses linkage distance.



WCSS

- Within-Cluster-Sum-of-Squares (WCSS)- Implicit **objective function in k-Means** measures sum of distances of observations from their cluster centroids.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$



Yi is centroid for observation Xi.

- Given that k-Means has no in-built preference for right number of clusters, following are some of the common ways k can be selected:
  - Domain Knowledge
  - Rule of Thumb
  - Elbow-Method using WCSS
  - Cluster Quality using Silhouette Coefficient

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i$$

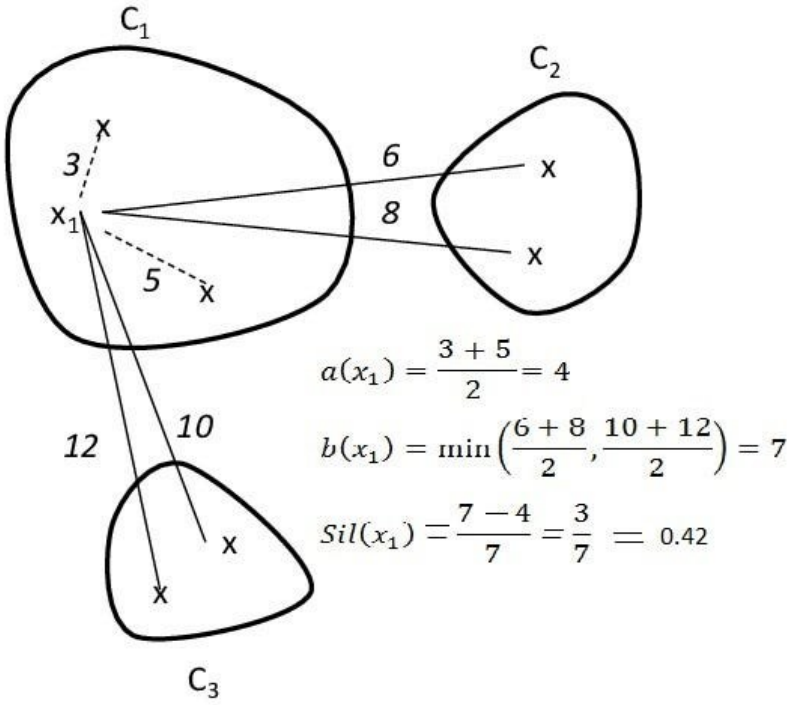
$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1}$$

Clustering Quality

- Silhouette Score
- Davies-Bouldin score
- Calinski-Harabasz score

Silhouette Score

- A metric for evaluating the quality of clustering results.
- Offers a quantitative measure to assess the appropriateness of clustering algorithms and aids in identifying the optimal number of clusters.
- The best value is 1 and
- Worst value is -1.
- Values near 0 indicate overlapping clusters.
- Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.



$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1\ldots nc, i\neq j}\left(R_{ij}\right), \; i = 1\ldots n_c$$

Davies-Bouldin score

- average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances.
- clusters which are farther apart and less dispersed will result in a better score.
- The minimum score is zero, with lower values indicating better clustering.

Calinski-Harabasz score

- It is also known as the Variance Ratio Criterion.

- The score is defined as ratio of the sum of between-cluster dispersion and of within-cluster dispersion.

$$k = \text{Number of clusters}$$
$$n_q = \text{Number of points in cluster } q$$
$$c_q = \text{Cluster center of cluster } q$$
$$n_E = \text{Number of data points}$$
$$c_E = \text{Cluster center of all points}$$
$$B = \sum_{q \in k} n_q (c_q - c_E)(c_q - c_E)^T$$
$$W = \sum_{q \in k} \sum_{x \in \text{cluster } q} (x - c_q)(x - c_q)^T$$
$$\text{CH} = \frac{B}{W} \times \frac{n_E - k}{k - 1}$$

Conclusions

Clustering Type	Algorithm	Description	Example Use Cases
Centroid-based Clustering	K-Means	Divides data into 'k' clusters by minimizing the distance between data points and the centroids of their respective clusters. It assigns each data point to the cluster with the nearest mean. It's effective for spherical clusters of similar sizes.	Segmenting customer data based on purchasing behavior to identify different customer segments for targeted marketing campaigns.
	Means Shift	Identifies clusters by shifting data points towards the mode of the kernel density estimate. It iteratively moves each data point to the local maximum of the density function until convergence, determining cluster centroids dynamically. Ideal for identifying clusters with arbitrary shapes and densities.	Identifying regions of interest in an image based on color and texture features for image recognition tasks.
Density-based Clustering	DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	Groups together data points that are closely packed together and marks points in low-density regions as outliers. It clusters dense regions of points separated by regions of lower density, without requiring the number of clusters as input. Suitable for datasets with varying cluster densities and shapes.	Identifying anomalous activities in network traffic data for cybersecurity purposes.
	OPTICS (Ordering Points To Identify the Clustering Structure)	Orders data points based on their density and spatial connectivity, allowing the discovery of clusters of varying density. It constructs a reachability plot to identify clusters and can handle clusters of arbitrary shapes and sizes. Effective for large datasets with noise and varying densities.	Clustering geographical data to identify regions of high and low population density for urban planning and resource allocation.
Distribution-based Clustering	Gaussian Mixture Models (GMM)	Assumes that the data is generated from a mixture of several Gaussian distributions and assigns probabilities to data points belonging to each cluster. It models clusters as ellipsoids in the feature space and assigns soft membership to points. Suitable for datasets with overlapping clusters.	Analyzing gene expression data to identify patterns in gene expression across different conditions or diseases.
	Expectation-Maximization (EM) Algorithm	A statistical method used to estimate parameters of a mixture model. It iteratively estimates the parameters of the model, such as means and covariances, by maximizing the likelihood function of the observed data, often used with GMMs for clustering. Effective for datasets with complex data distributions.	Identifying clusters of user behavior in web usage logs to personalize website content and improve user experience.
Hierarchical Clustering	BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)	Constructs a hierarchical clustering structure by creating a tree of clusters, known as the Clustering Feature Tree (CF Tree). It incrementally processes data points and merges them into clusters while maintaining a compact and efficient representation. Ideal for handling large datasets efficiently.	Clustering news articles based on their content to organize them hierarchically for easier navigation and recommendation systems.