Unsupervised Learning (Clustering)

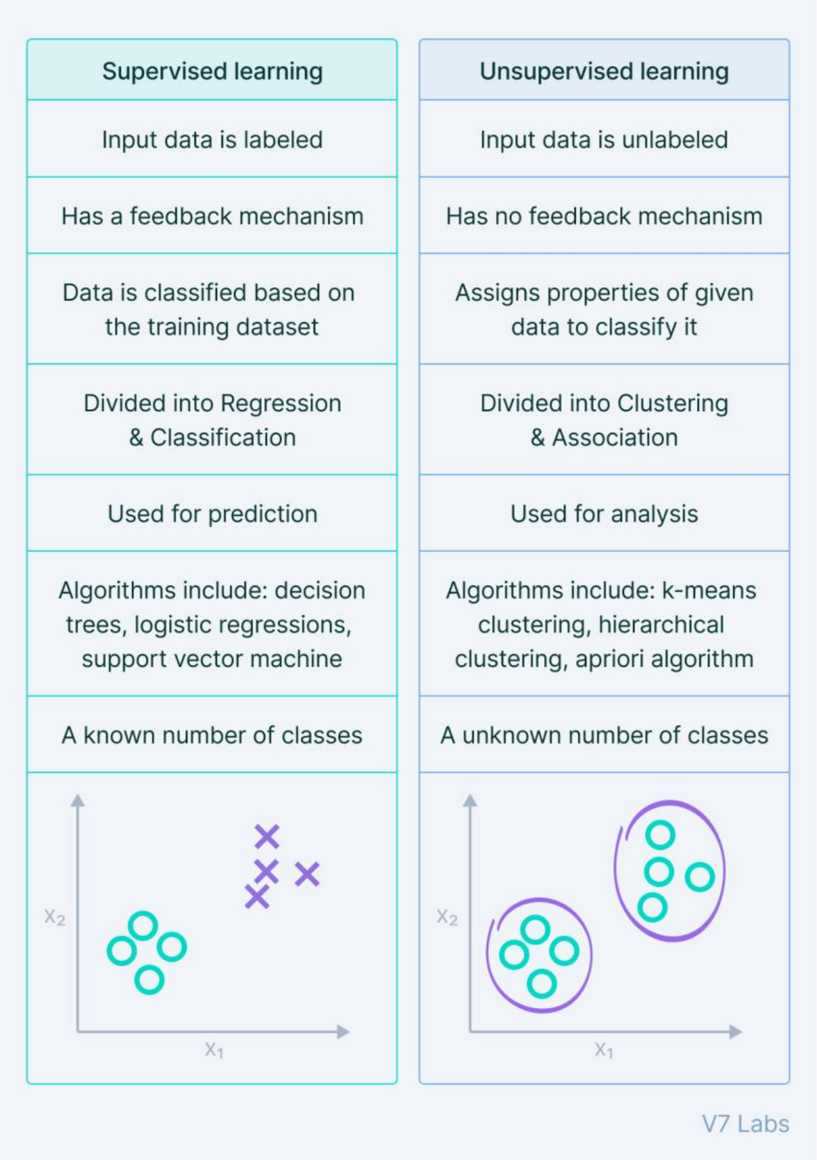
Penyusun Modul: Chairul Aulia

Editor: Citra Chairunnisa

Unsupervised Learning

Unsupervised learning is the training of a machine using information that is **neither classified nor labeled** and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to **similarities, patterns, and differences**.

Unlike supervised learning, no teacher is provided that means **no training** will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.



Supervised learning	Unsupervised learning
Input data is labeled	Input data is unlabeled
Has a feedback mechanism	Has no feedback mechanism
Data is classified based on the training dataset	Assigns properties of given data to classify it
Divided into Regression & Classification	Divided into Clustering & Association
Used for prediction	Used for analysis
Algorithms include: decision trees, logistic regressions, support vector machine	Algorithms include: k-means clustering, hierarchical clustering, apriori algorithm
A known number of classes	A unknown number of classes

V7 Labs

Types of Unsupervised Learning

- **Clustering** is the type of Unsupervised Learning where we find hidden patterns in the data based on their **similarities or differences**.
- **Association** is the kind of Unsupervised Learning where we can find the **relationship** of one data item to another data item or discover the **rules**, such as people that buy X also tend to buy Y

Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points are more similar in the same groups and dissimilar in the other groups. It is basically a collection of objects on the basis of **similarity and dissimilarity** between them.

Applications of Clustering

- **Marketing**: It can be used to characterize & discover customer segments for marketing purposes.
- **Biology**: It can be used for classification among different species of plants and animals.
- **Libraries**: It is used in clustering different books on the basis of topics and information.
- **Insurance**: It is used to acknowledge the customers, their policies and identifying the frauds.
- **City Planning**: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
- **Earthquake studies**: By learning the earthquake-affected areas we can determine the dangerous zones.
- **Search Engines**: The search result appears based on the closest object to the search query.

Types of Clustering Methods

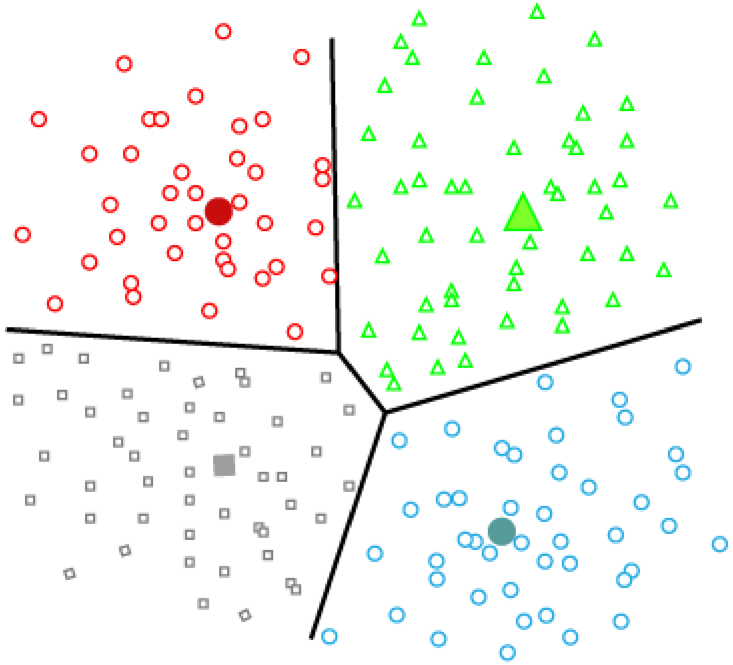
The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and Soft **Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

1. Partitioning Clustering
2. Density-Based Clustering
3. Distribution Model-Based Clustering
4. Hierarchical Clustering
5. Fuzzy Clustering

Partitioning Clustering

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method.

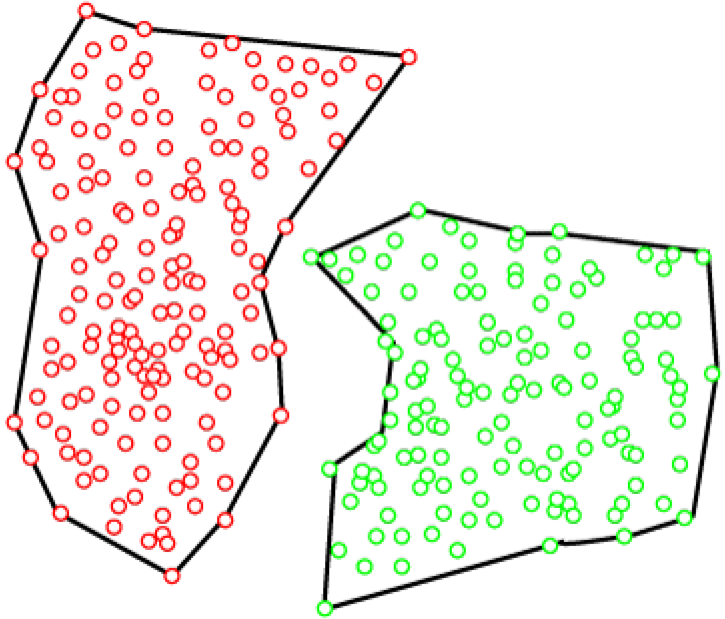
This is basically one of the iterative clustering algorithms in which the clusters are formed by the **closeness of data points to the centroid** of clusters. Here, centroid is formed such that the distance of data points is minimum with the center and the number of clusters to be created is specified.



Algorithm examples: K-Means

Density-Based Clustering

The density-based clustering method **connects the highly-dense areas** into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.



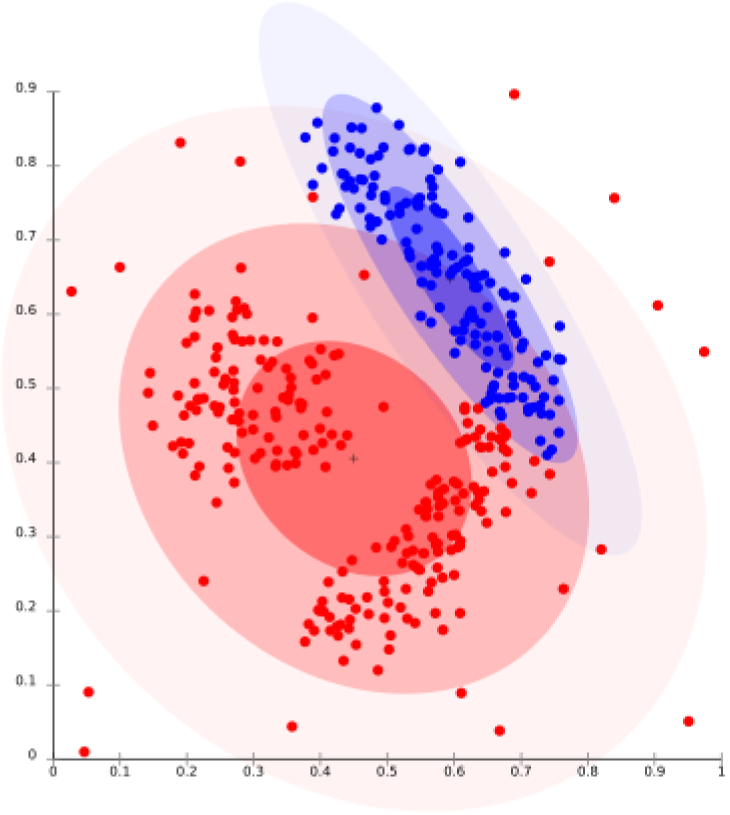
These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

Algorithm examples: DBSCAN, OPTICS

Distribution Model-Based Clustering

It is a clustering model in which we will fit the data on the probability that **how it may belong to the particular distribution**. The grouping is done by assuming some distributions commonly Gaussian Distribution.
Distribution-based clustering produces clusters that assume concisely defined mathematical models underlying the data, a rather strong assumption for some data distributions.

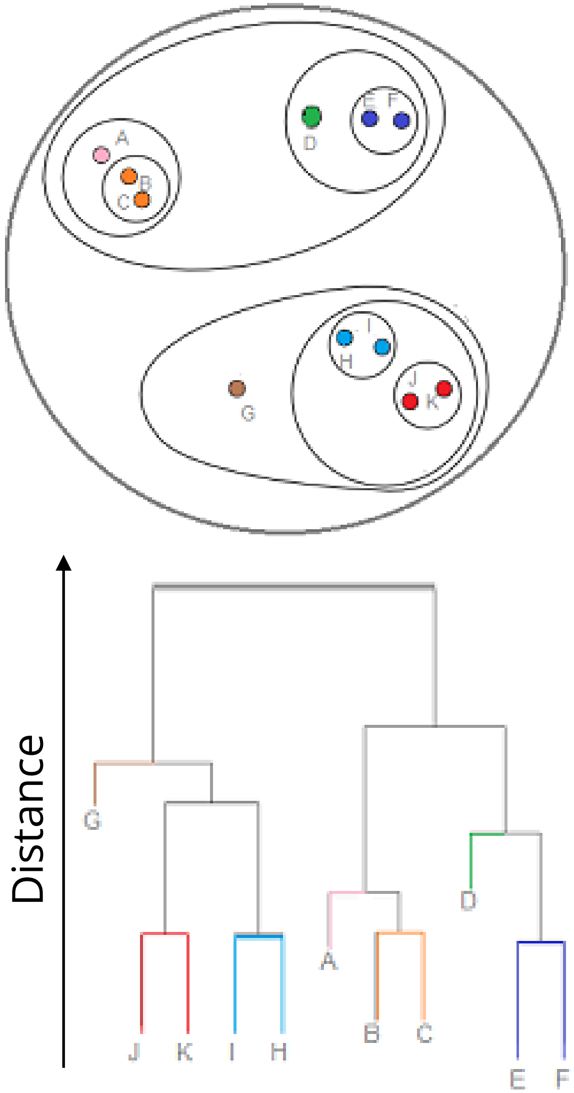
Algorithm example: Gaussian Mixture Model (GMM)



Hierarchical Clustering

It can be used as an alternative for the partitioned clustering as there is no requirement of specifying number of clusters. It is also known as connectivity-based methods. In this method, it provides us **with the hierarchy of the clusters that merge after a certain distance**. After the clustering is done, the result will be a tree-based representation of data points (Dendogram), divided into clusters.
Number of clusters can be selected by cutting the tree at the correct level.

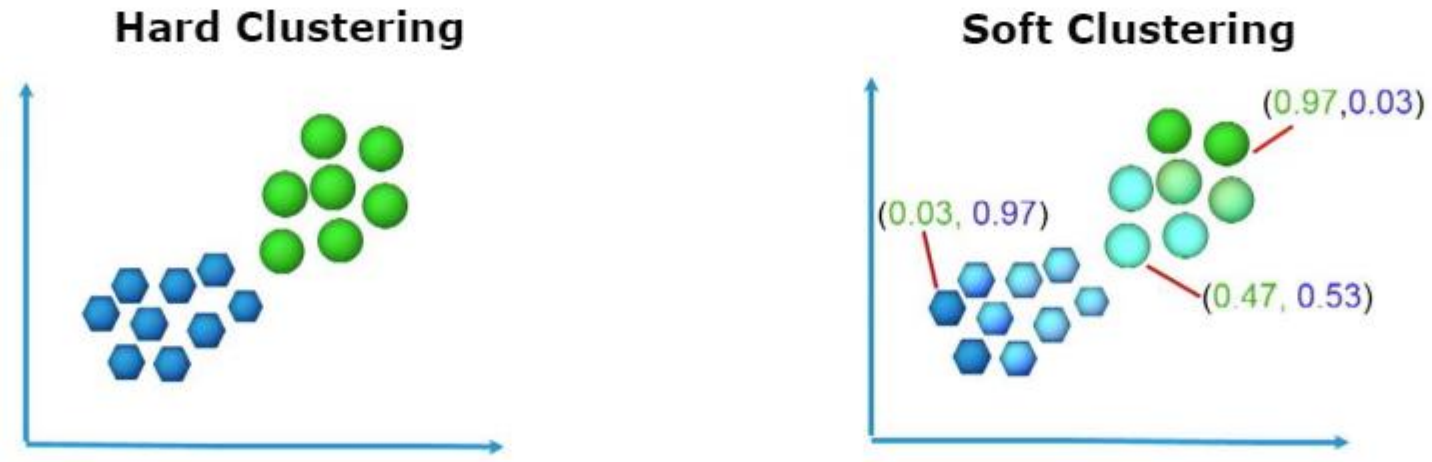
Algorithm examples: BIRCH, Affinity propagation, Agglomerative



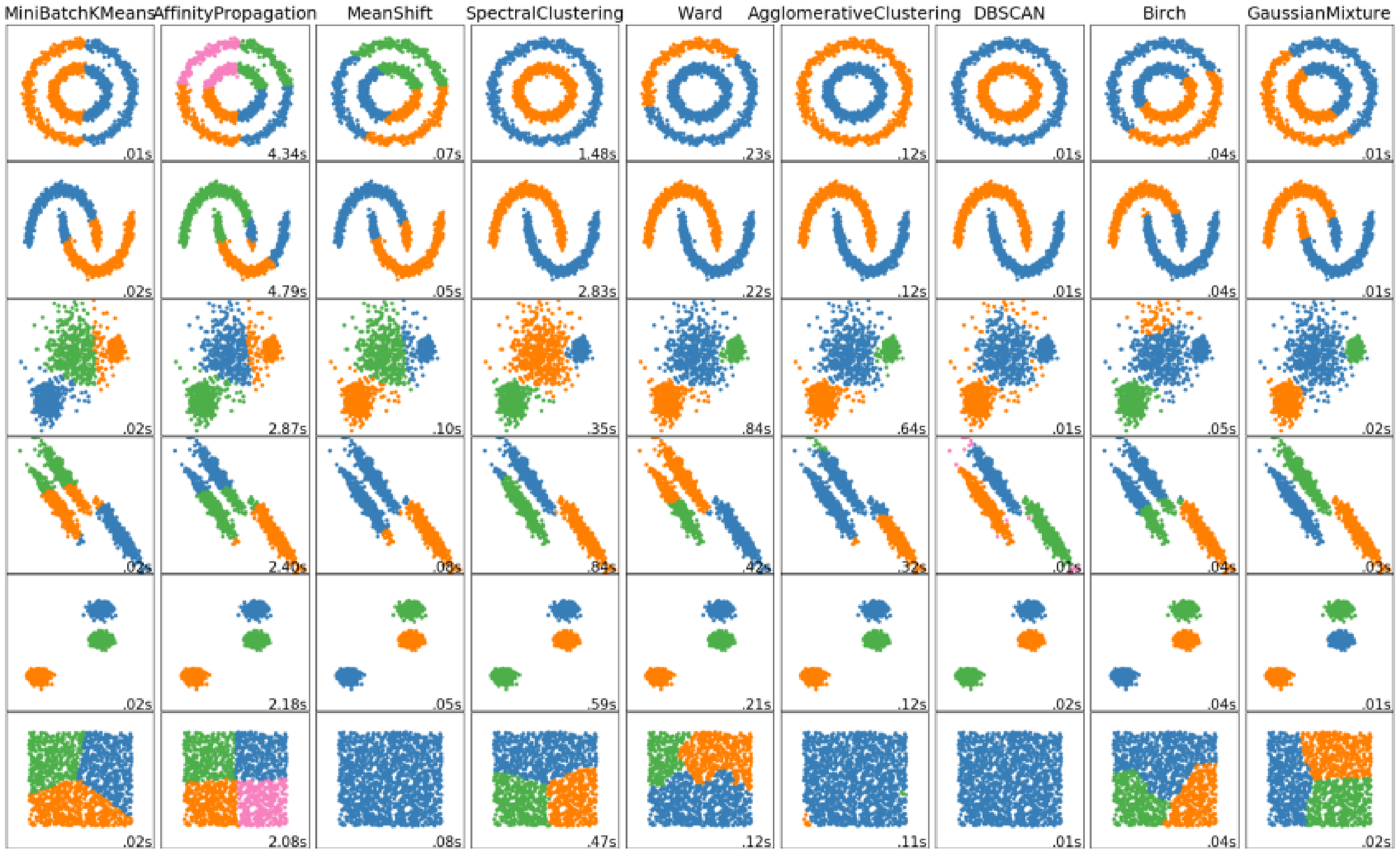
Dendogram

Fuzzy Clustering

It belongs to a branch of **soft method clustering** techniques, whereas all the above-mentioned clustering techniques belong to hard method clustering techniques. In this clustering technique, points close to the center may be a part of the other cluster to a higher degree than points at the same cluster's edge. The probability of a point belonging to a given cluster is a value that lies between 0 to 1.
Algorithm examples: Fuzzy C-Means



Different Algorithms on Different Kinds of Data



https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

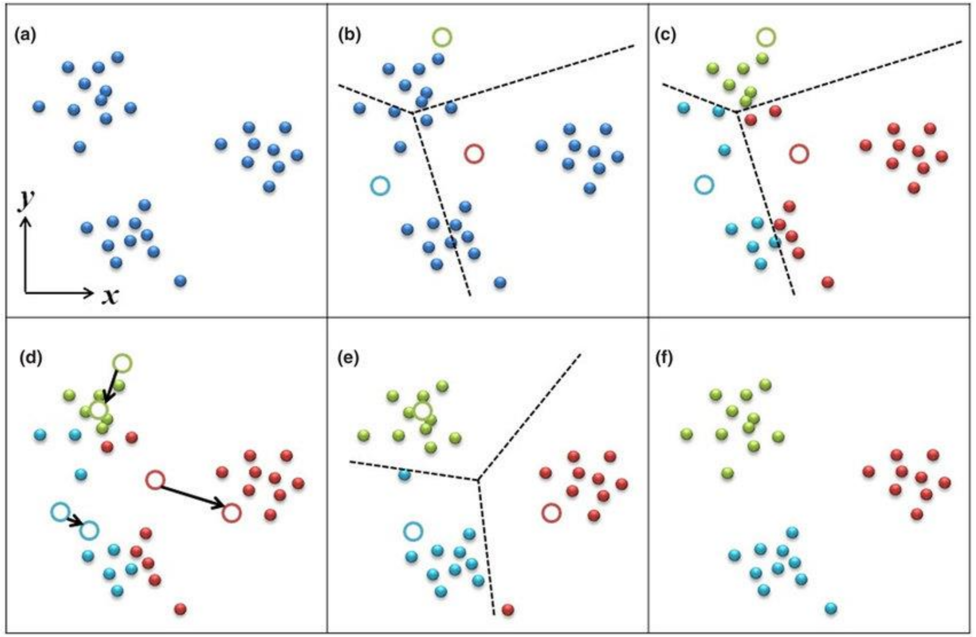
Please note that you don’t have to remember all the algorithms, just know that there are multiple algorithms with different use cases, so you can just check them later and then compare to see which one is more suitable for your problem?

K-Means Algorithm

The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The end results of the K-means clustering algorithm would be:

- Centroids of the number of clusters, which were identified (denoted as K).
- Labels for the training data.

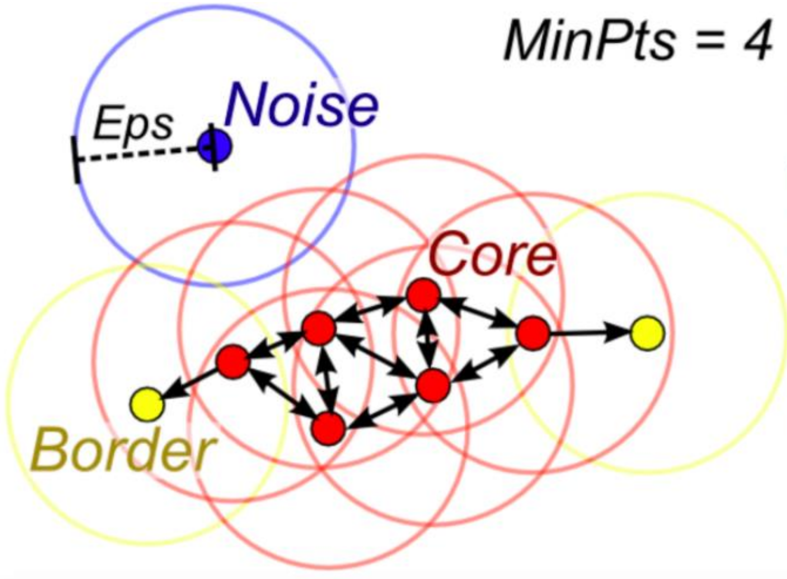
Challenge: finding the number of clusters (K)
There is no one-size-fits-all solution to this problem. Yet, specific performance measures might help, i.e. **elbow method**. The elbow method is based on measuring the inertia. Inertia is sum of squared distances of samples to their closest cluster center



DBSCAN Algorithm

It groups ‘densely grouped’ data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

DBSCAN requires only two parameters: **epsilon** and **minPoints**. Epsilon is the radius of the circle to be created around each data point to check the density and minPoints is the minimum number of data points required inside that circle for that data point to be classified as a Core point.



Red: Core Points

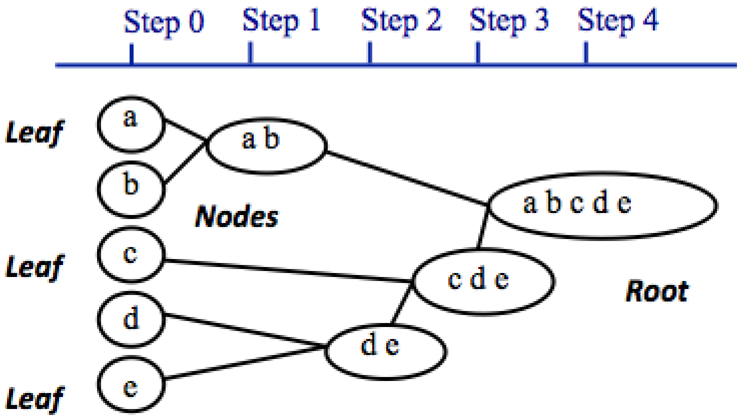
Yellow: Border points. Still part of the cluster because it’s within epsilon of a core point, but not does not meet the min_points criteria

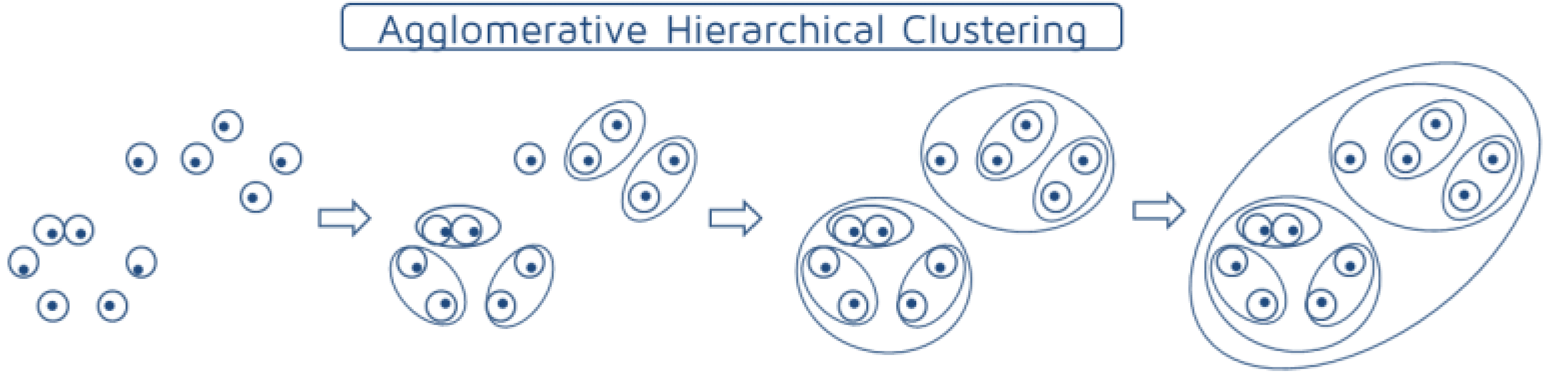
Blue: Noise point. Not assigned to a cluster

https://medium.com/geekculture/dbscan-clustering-algorithm-simply-explained-993c195d1f63

Agglomerative Algorithm

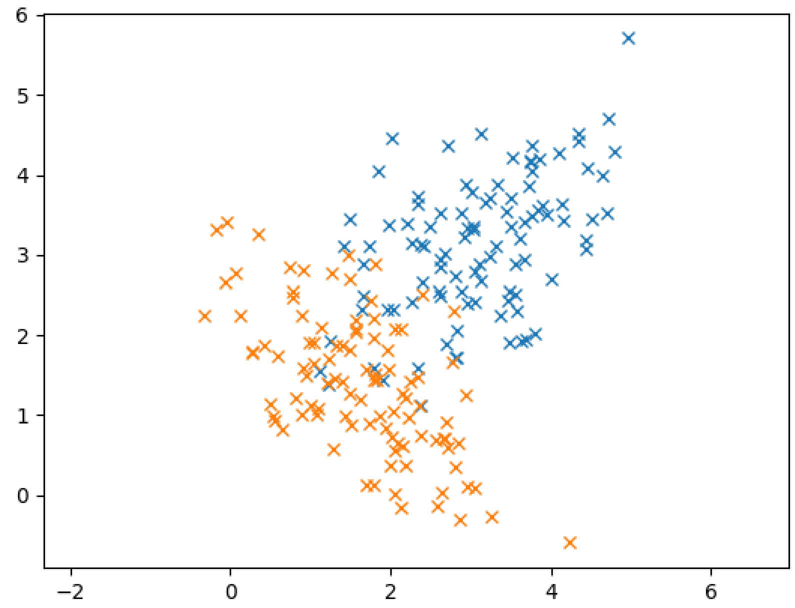
Agglomerative clustering works in a “bottom-up” manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root). The steps can be:





Gaussian Mixture Algorithm

Consider there is a case where we have overlap in our data and K-Means doesn’t seem to work well. Suppose we know the distribution and it is Gaussian, we can use this info and use Gaussian Mixture algorithm.
It turns out that many dataset distributions are actually Gaussian! There is a famous theorem in statistics called the **Central Limit Theorem** that states that as we collect more and more samples from a dataset, they tend to resemble a Gaussian, even if the original distribution is not Gaussian! This makes Gaussian very powerful and versatile)



Exercise

Let’s implement several clustering algorithms with sklearn

Hint:

```
from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering
from sklearn.mixture import GaussianMixture
```

Generate dataset with `make_blobs()`

Exercise

Let's cluster the mall customers based on their annual income and spending score. Use K-Means and choose number of cluster with elbow method

Use **Mall_customers** datasets