Machine Learning

Session 3 - Clustering

✉ hadrien.salem@centralelille.fr

introduction-to-data-science

Introduction

What did we do last time?

Course outline

Machine learning course

Session 1: Regression

Session 2: Supervised classification

Session 3: Clustering

Session 4: Decision trees and ensemble methods

Session 5: Introduction to neural networks

Session 6: Advanced neural networks

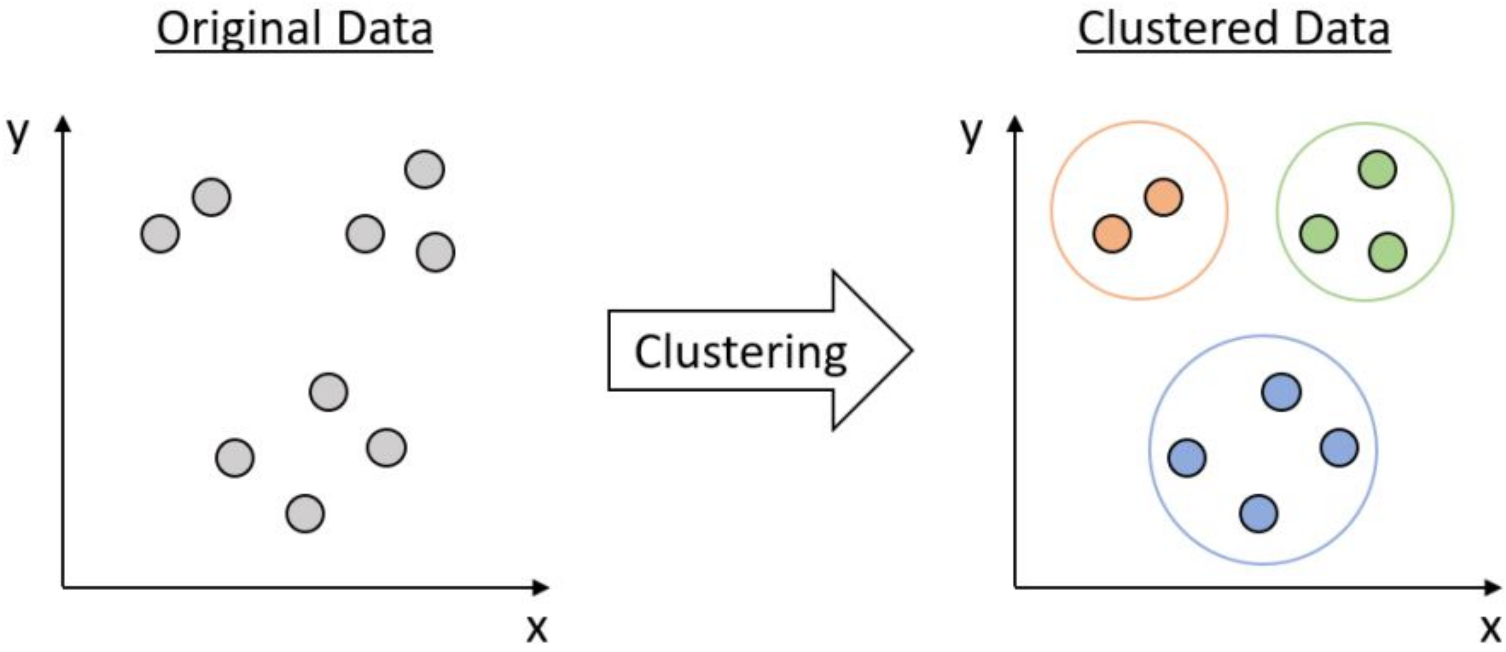
Session 7: Introduction to reinforcement learning

Session 8: Reading science papers

↓

Project

What is clustering?



Clustering is a **machine learning task** that consists in **finding groups of similar data points**

Clustering is a **task** consisting in **grouping similar data points** together without using predefined labels.

Clustering is equivalent to unsupervised classification

Data points in a group (or cluster) should be more similar to each other than to those in other groups

What are some concrete applications of clustering?

What are some concrete applications of clustering?

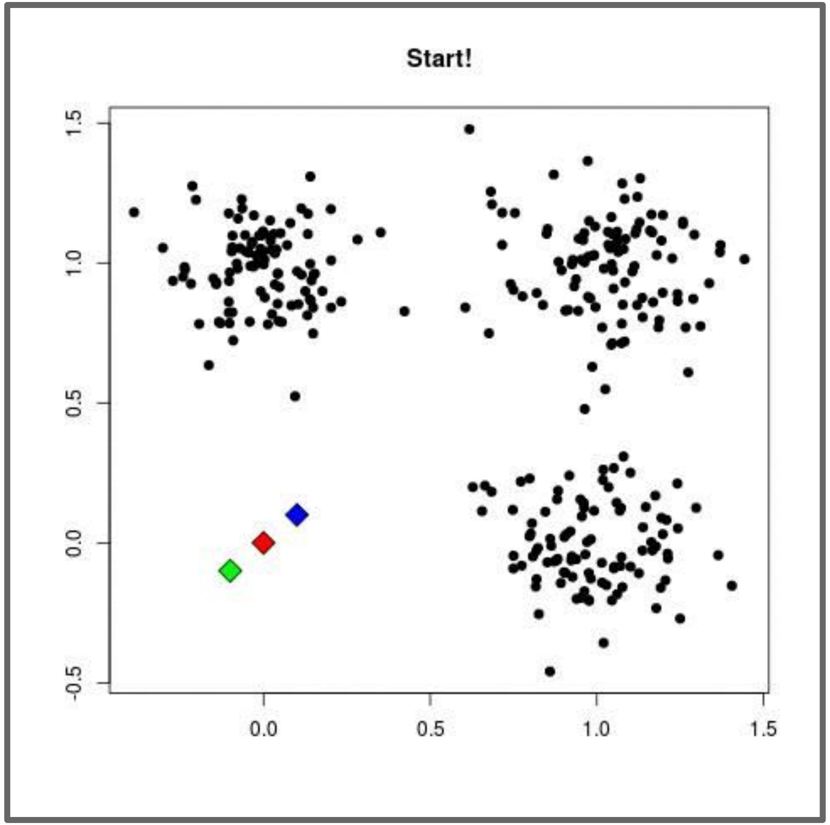
Clustering can be used for…

- Creating groups of similar customers (Market / Customer segmentation)
- Anomaly Detection
- Image segmentation
- Organizing collections of documents
- etc.

Clustering methods

Clustering algorithms family #1
Partitional clustering

Partitional clustering divides the data into classes based on their similarity
It requires specifying the number of clusters beforehand

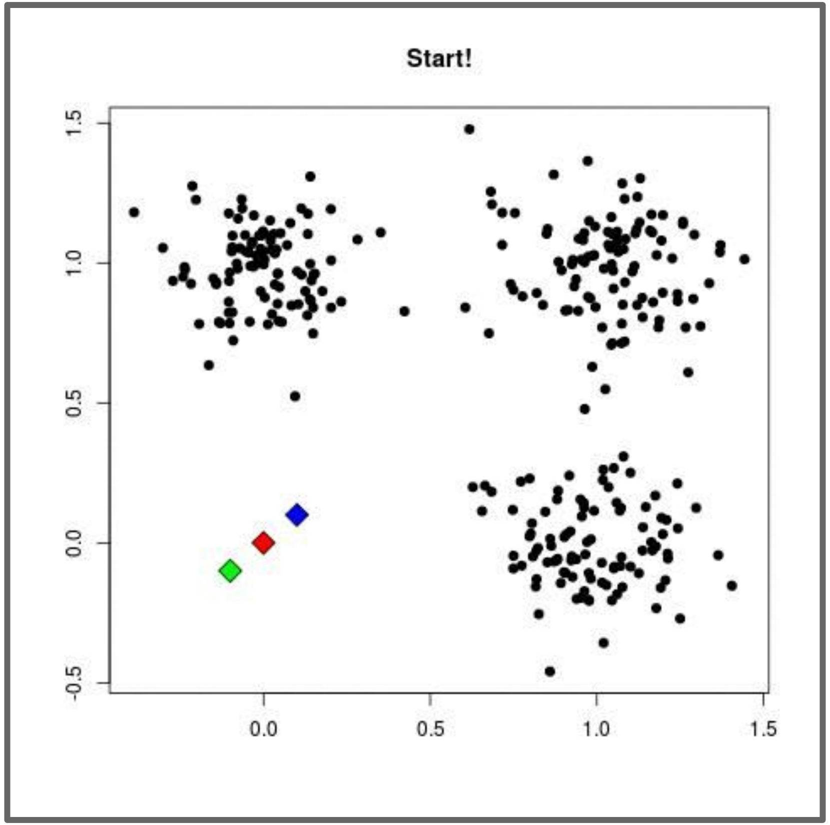


Procedure

1. Choose a value for K (number of clusters)
2. Select random cluster centers
3. Assign data points to clusters
4. Update cluster centers (mean of the cluster)
5. Repeat steps 3 and 4 until convergence

K-medoids is a variation of K-means where the cluster centers are actual data points.

K-means is one of the simplest clustering algorithms

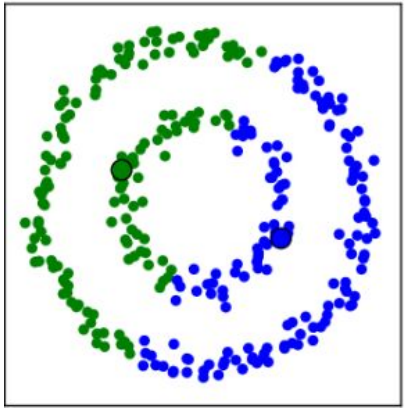


Strengths

- Simple and fast
- K-medoids is more robust to outliers (no mean)

Weaknesses

- Choosing K can be difficult
- Computing the mean is sensitive to outliers
- Clusters are assumed to be spherical and of similar sizes

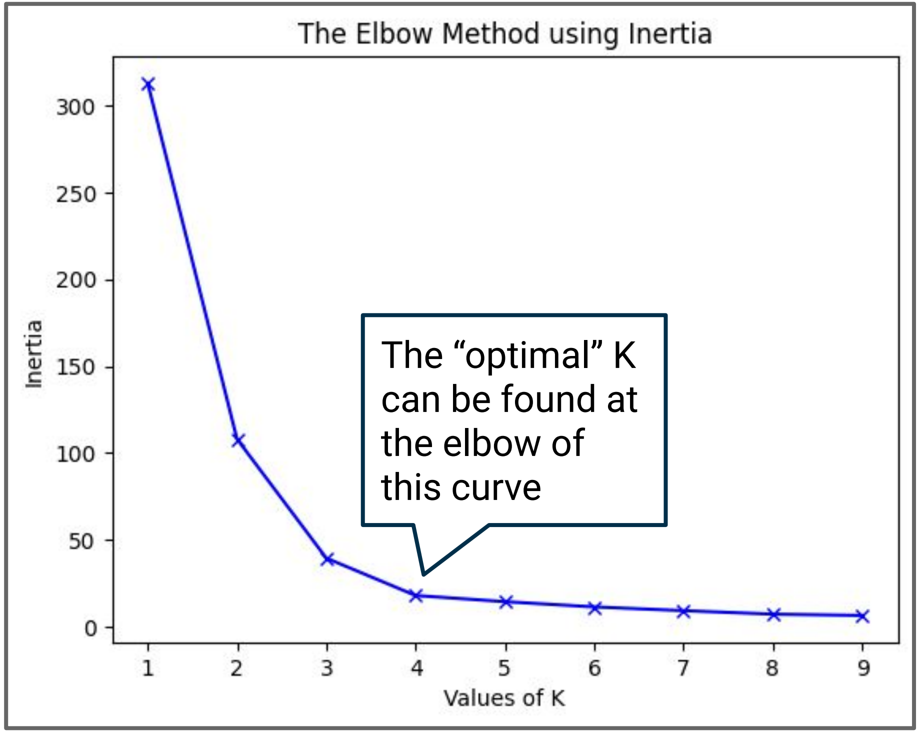


Strengths and weaknesses of K-means / K-medoids

Inertia
Sum of squared distances between data points and their assigned cluster center

Lower inertia means more compact clusters

In general, it is difficult to evaluate the performance of clustering algorithms without external labels



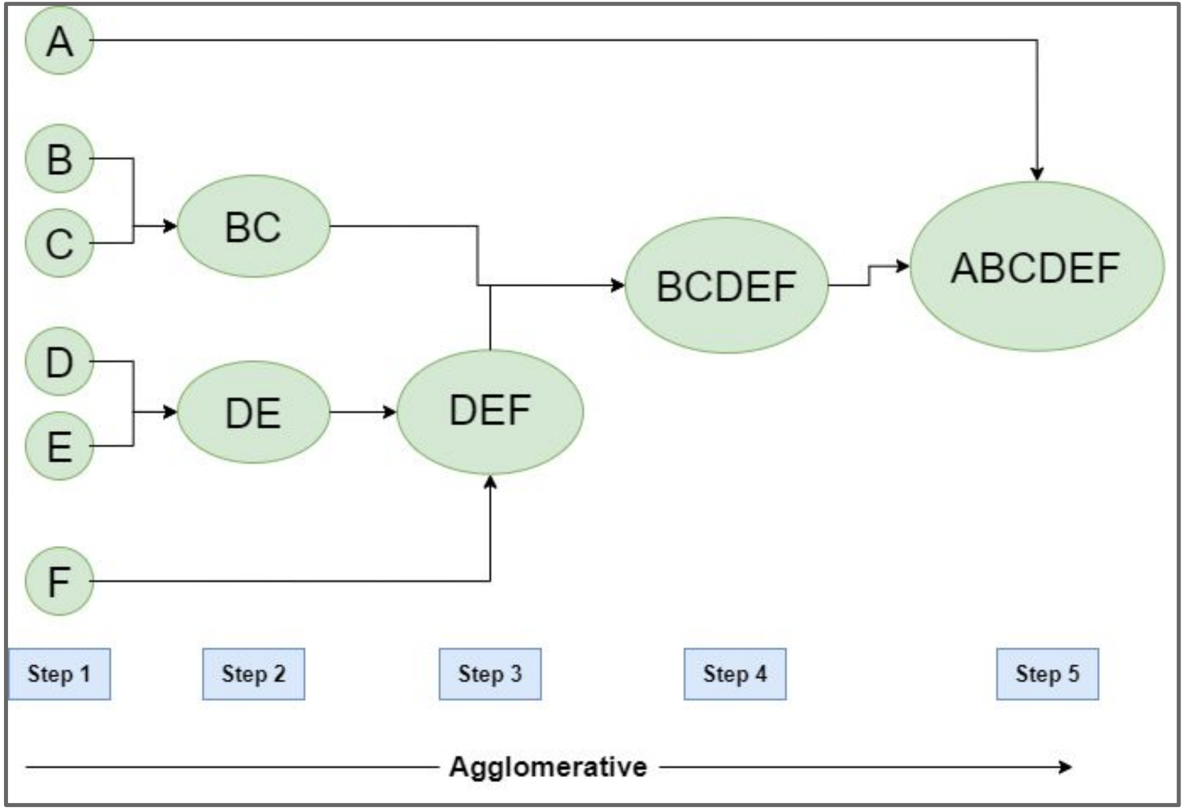
The “optimal” K can be found at the elbow of this curve

Choosing the value of K – The elbow method

Clustering algorithms family #2
Hierarchical clustering

Hierarchical clustering consists in building a hierarchy of clusters by merging or dividing clusters

It does not require specifying the number of clusters beforehand

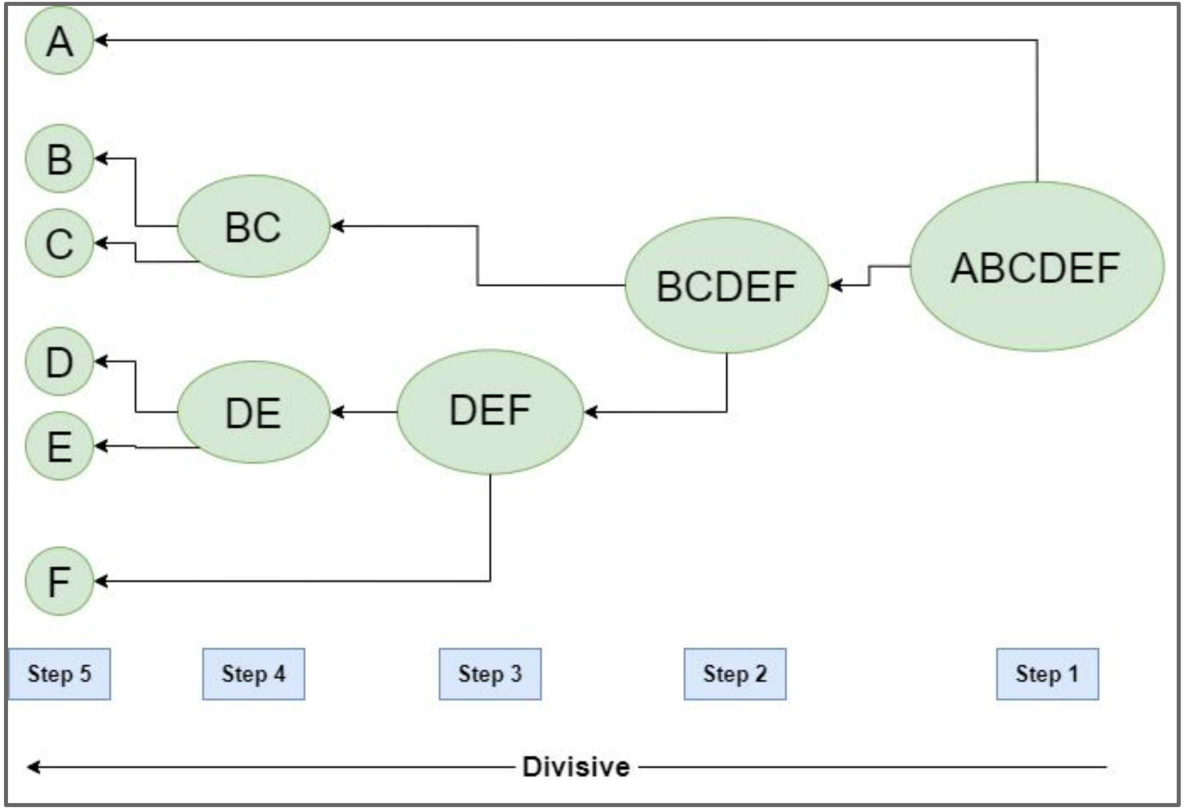


Agglomerative clustering

1. Each data point starts as a single cluster
2. At each step, the “closest” clusters are merged together
3. Repeat until all data points are grouped into a single cluster

The number of clusters can be selected by choosing at what step the algorithm is stopped

Image source



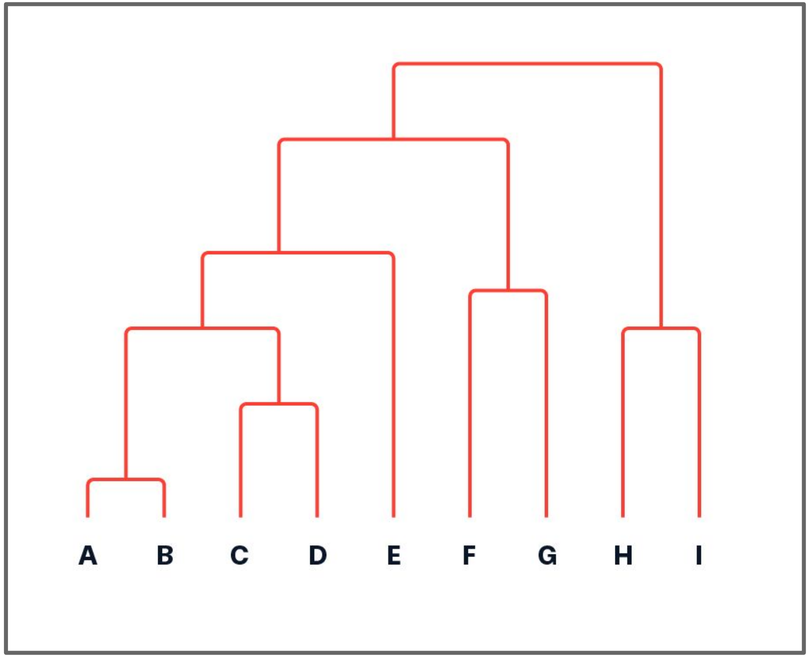
Agglomerative clustering

1. Start with a cluster containing all data points
2. At each step, the clusters are divided into two smaller clusters with a criteria to define
3. Repeat until all data points are single clusters

The number of clusters can be selected by choosing at what step the algorithm is stopped

Divisive clustering

Image source



Strengths

- Can display a dendrogram for visualization
- Flexibility with the number of clusters
- Can adapt to nonlinear decision boundaries

Weaknesses

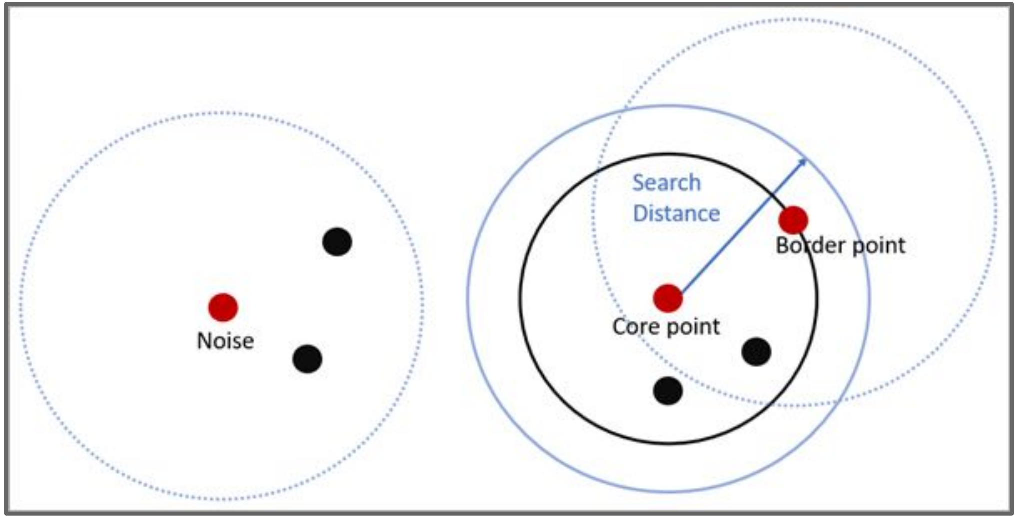
- Can be computationally expensive
- It can be difficult to choose the linking method
- Can be sensitive to outliers

Strengths and weaknesses of hierarchical clustering

Clustering algorithms family #3
Density-based clustering

Density-based clustering consists in grouping data points that are close to each other in areas of high density

It is well-suited for irregularly-shaped clusters



DBSCAN is an example of density-based clustering algorithm

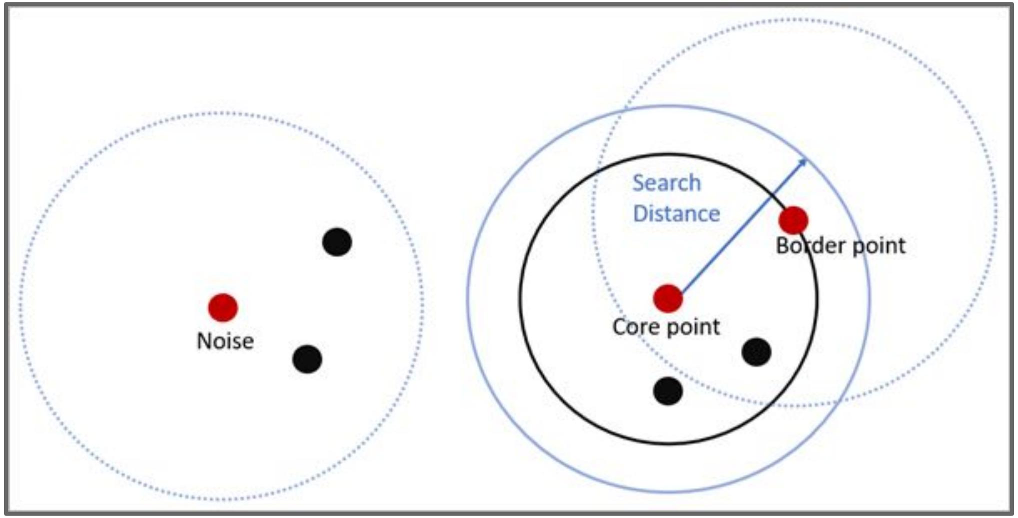
Core points are the points that have a certain number of other data points within a pre-defined search distance

Border points are within the search distance of a core point, but does not have enough neighbours to be considered a core point

Noise points do not match the previous conditions

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

Image source

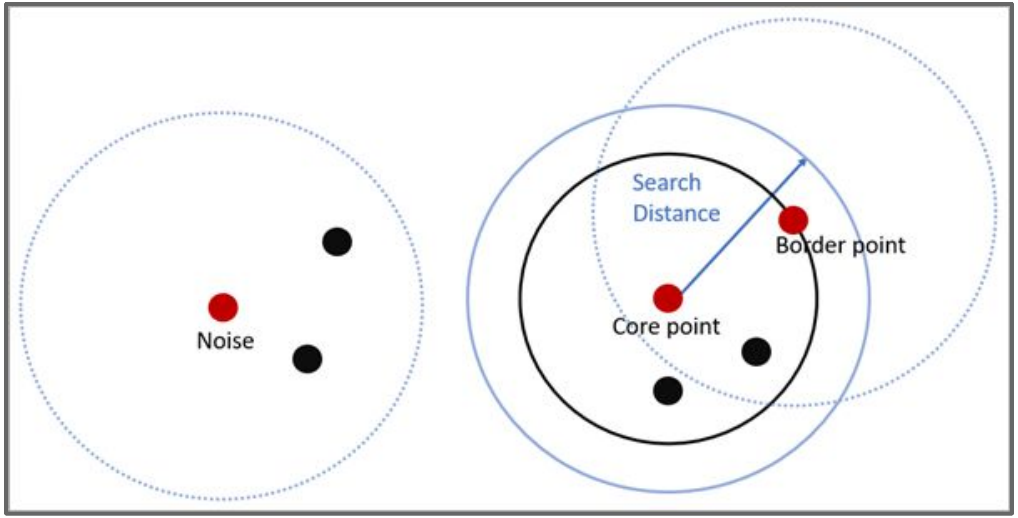


Procedure

1. Pick a point that has not been assigned to a cluster. If it is a core point, start a cluster around it. If not, mark it as noise
2. Expand the cluster by connecting all reachable points (core and border)
3. Repeat until all points are assigned to a cluster or marked as noise

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

Image source



Strengths

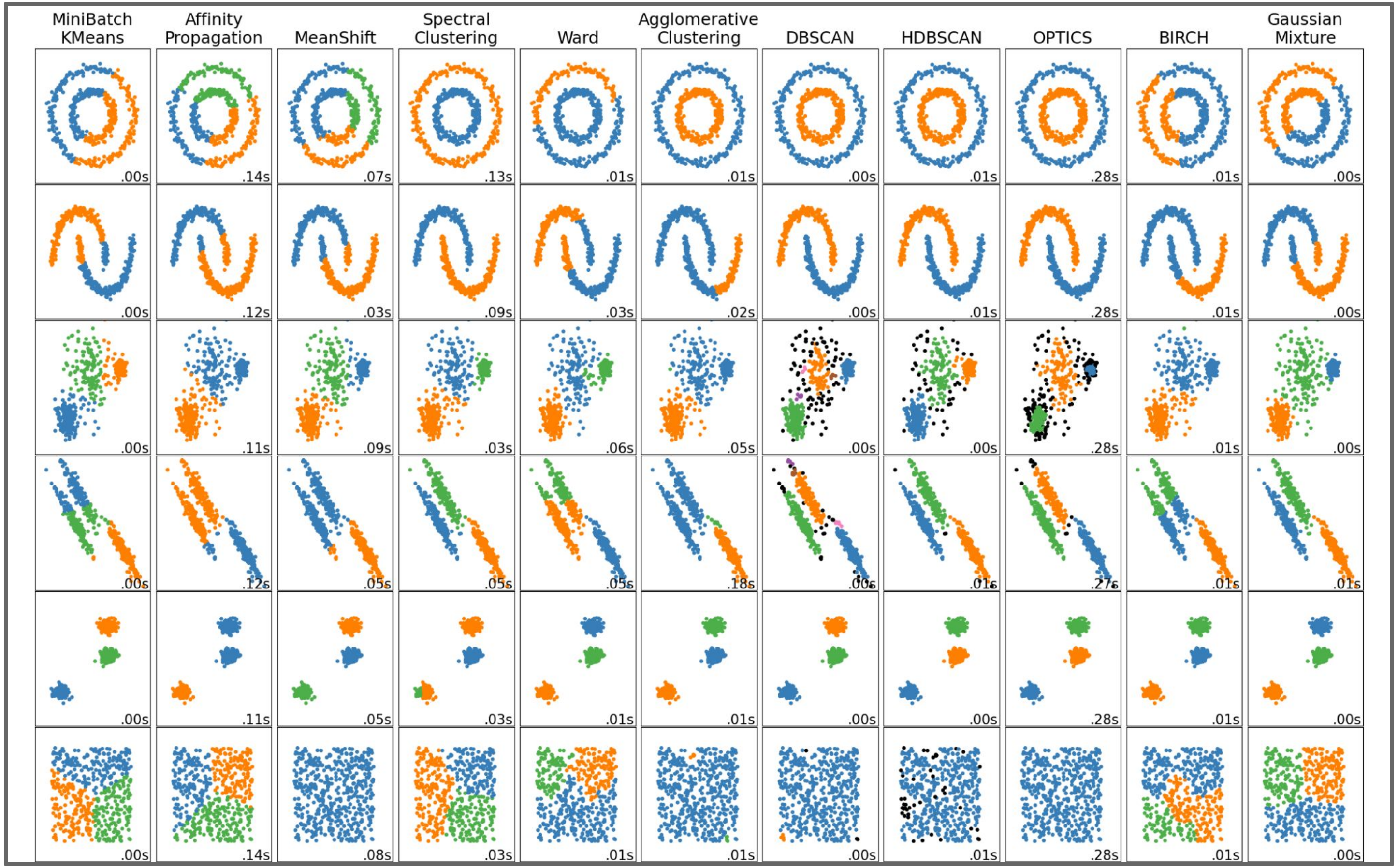
- Works with irregularly-shaped clusters
- Robust to noise and outliers
- No need to define the number of clusters

Weaknesses

- Not good with high dimensionality
- Struggles with varying densities
- Defining the search distance and number of neighbours can be difficult

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

Image source



All conventional clustering algorithms can be found on sk-learn with detailed documentation!

Practical work

The notebook contains all the necessary instructions

Debrief

Debrief

What did we learn today?

What could we have done better?

What are we doing next time?