Garanti BBVA Technology

Clustering

Sezgi Şener
Emre Tekin

October 2023

Garanti BBVA Technology

Agenda

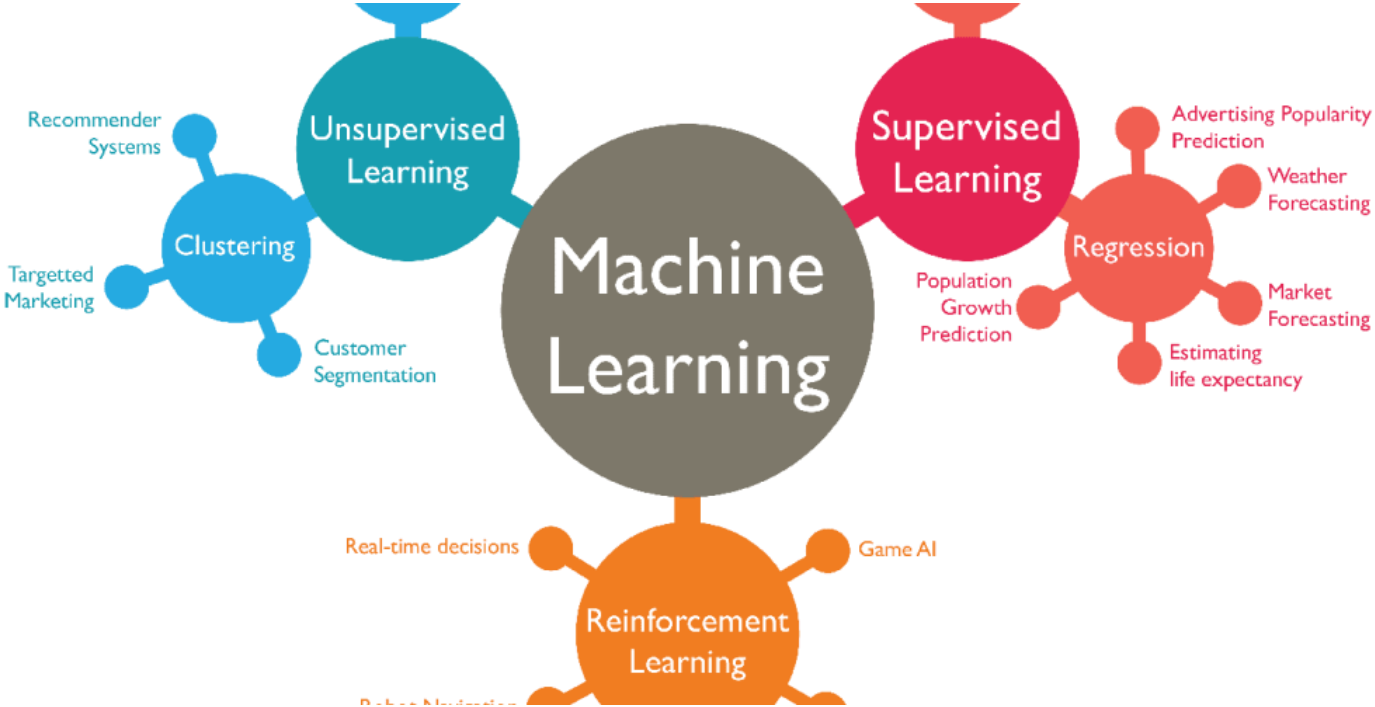
01 Overview

02 Partitioning Methods

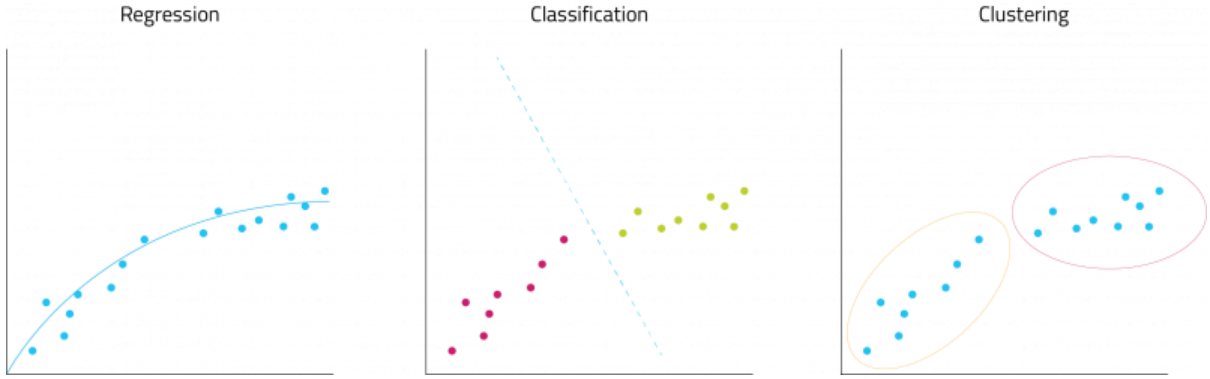
03 Hierarchical Methods

04 Density Based Methods

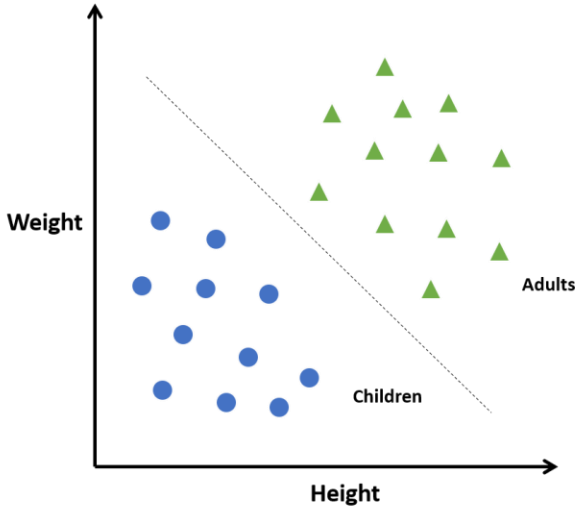
05 Interpreting Clustering Results

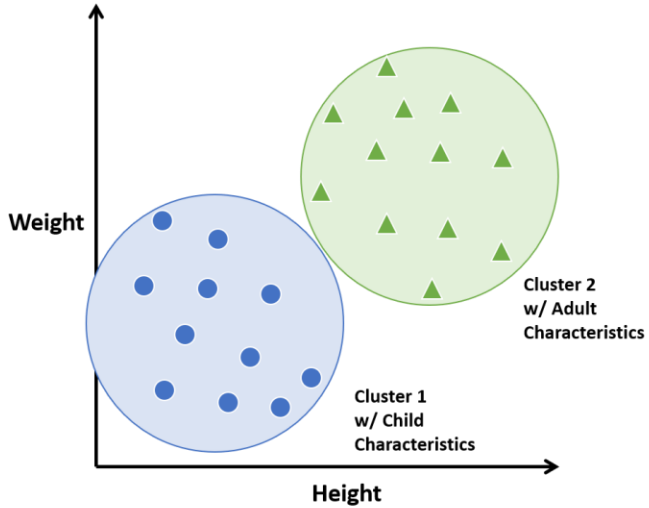


Machine Learning Problems



Classification vs Clustering





Sample Usage Areas

Determination of advertising campaigns by identifying customers with similar characteristics

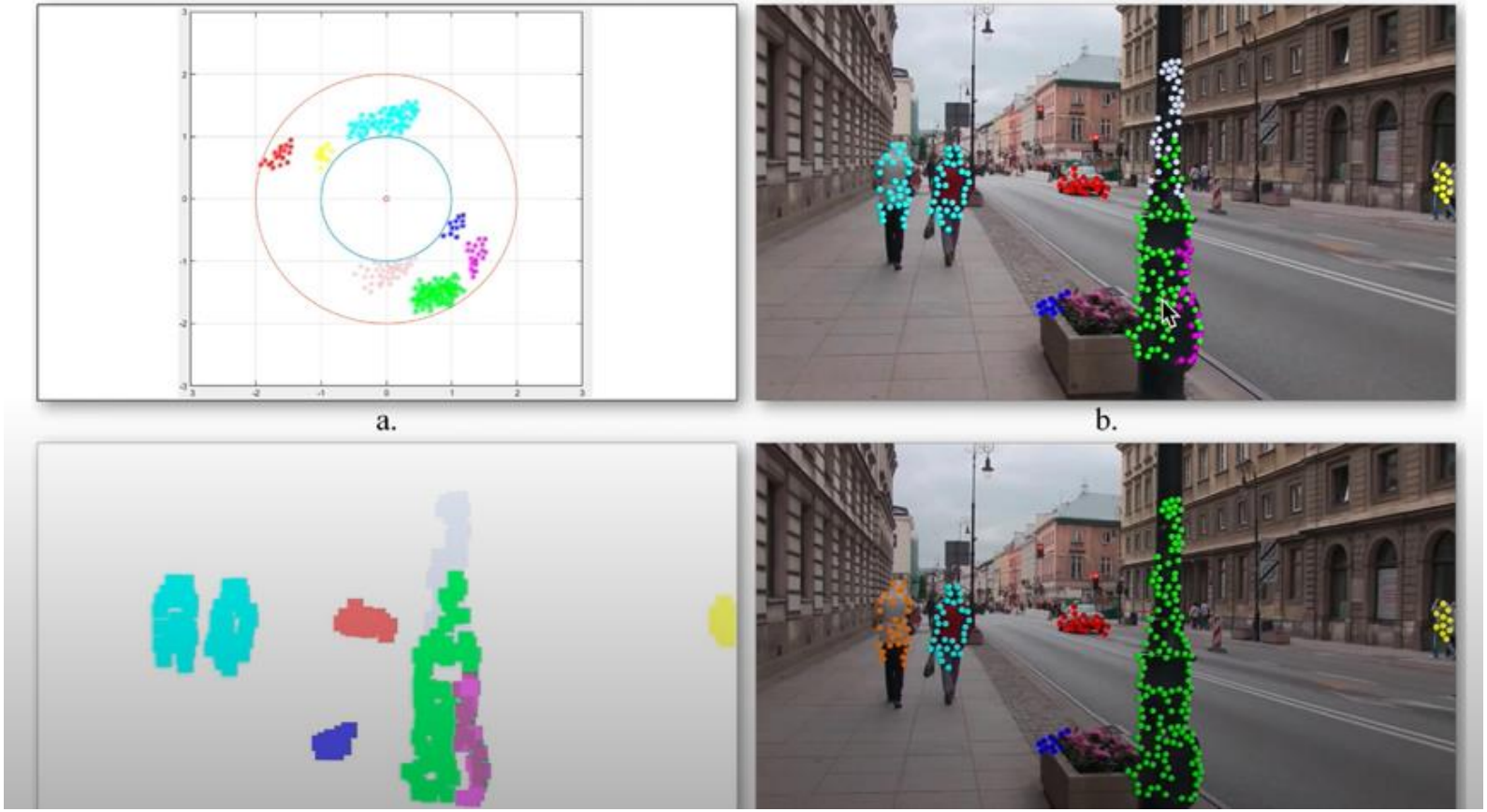
Grouping similar documents on the Internet

Revealing similar protein sequences

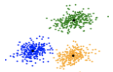
Detecting moving objects away from each other

Fraud Detection

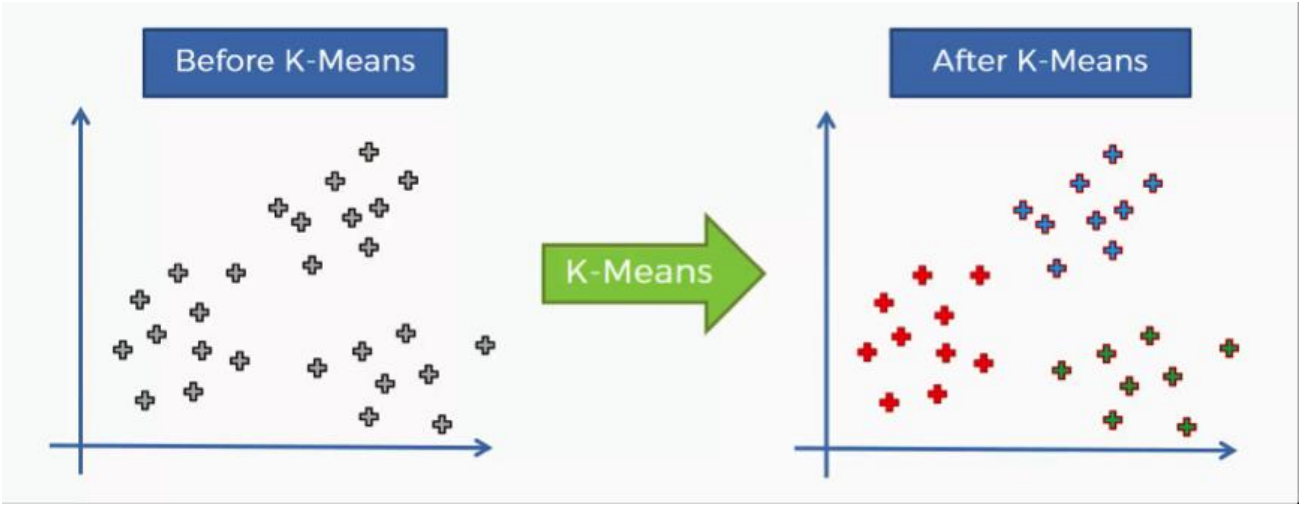
Completing missing values



Clustering Methods

- Partitioning Methods
  -  k means
  -  k medoids
- Hierarchical Methods
  -  Bottom-up
  -  Top-down
- Density-based Methods
  -  DBSCAN
  -  OPTICS

K Means Clustering



K Means Clustering

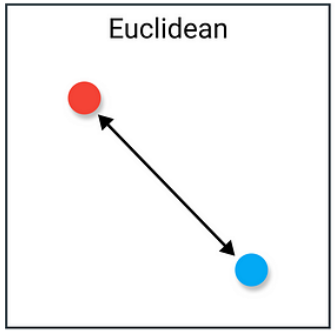
Randomly select central points

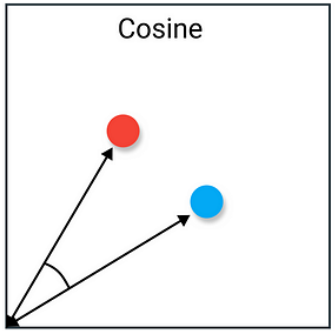
Assign each observation to the cluster with the nearest mean

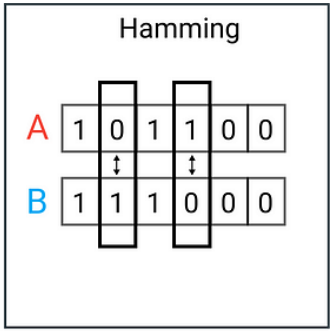
Recalculate means (centroids) for observations assigned to each cluster.

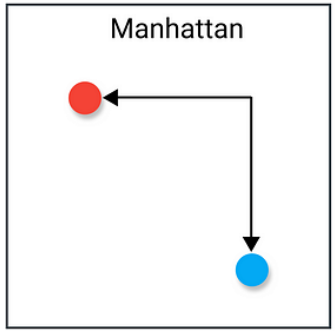
The algorithm has converged when the assignments no longer change.

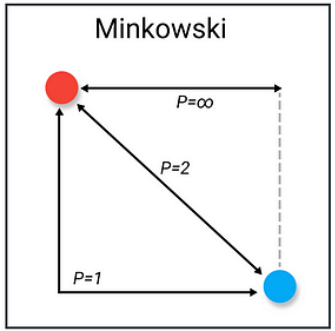
Distance Metrics

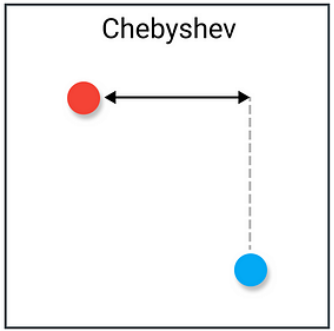


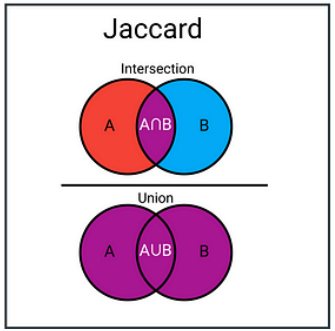


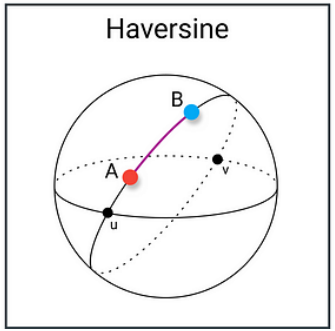


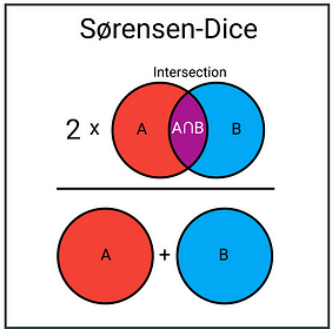










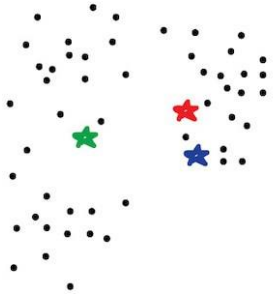


Animation

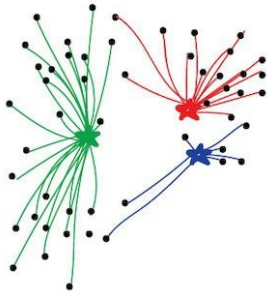
http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

PUT KEBAB KIOSKS IN THE OPTIMAL WAY

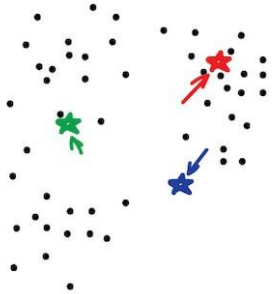
(also illustrating the K-means method)



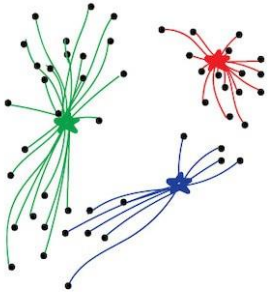
1. Put kebab kiosks in random places in city



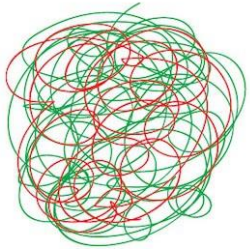
2. Watch how buyers choose the nearest one



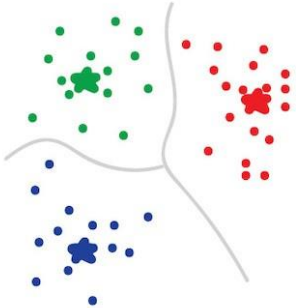
3. Move kiosks closer to the centers of their popularity



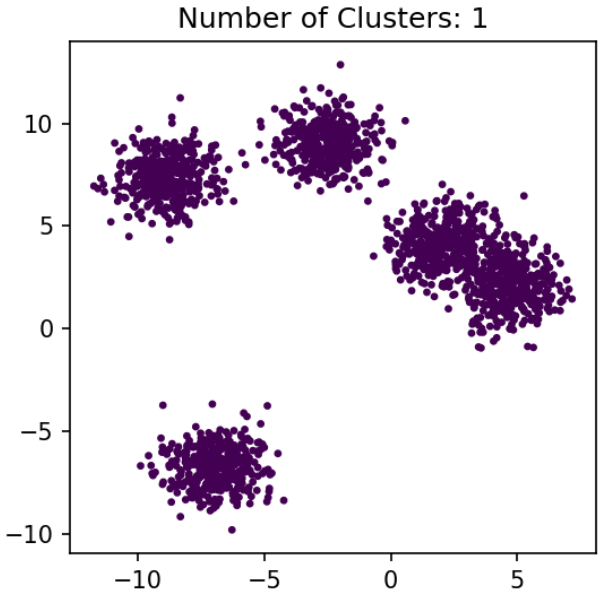
4. Watch and move again



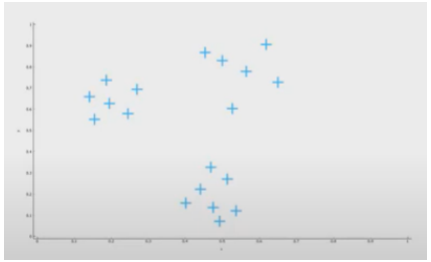
5. Repeat a million times

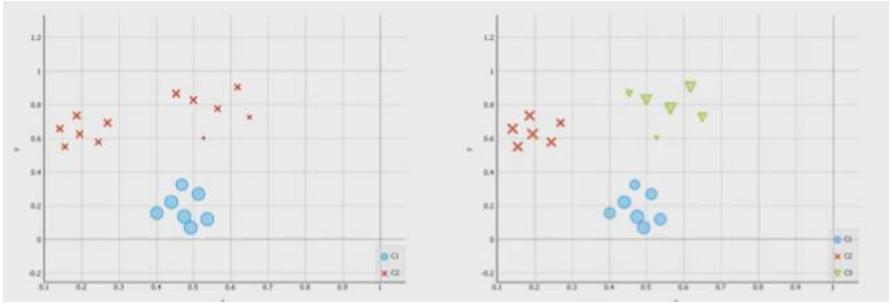


6. Done!
You're god of kebabs!



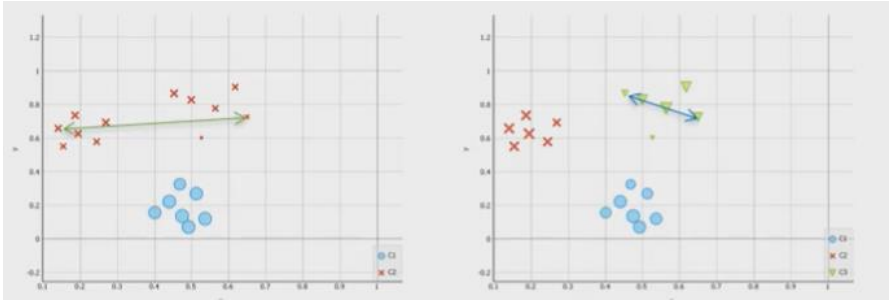
K = ? (Elbow method)



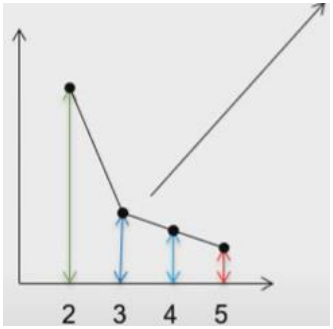


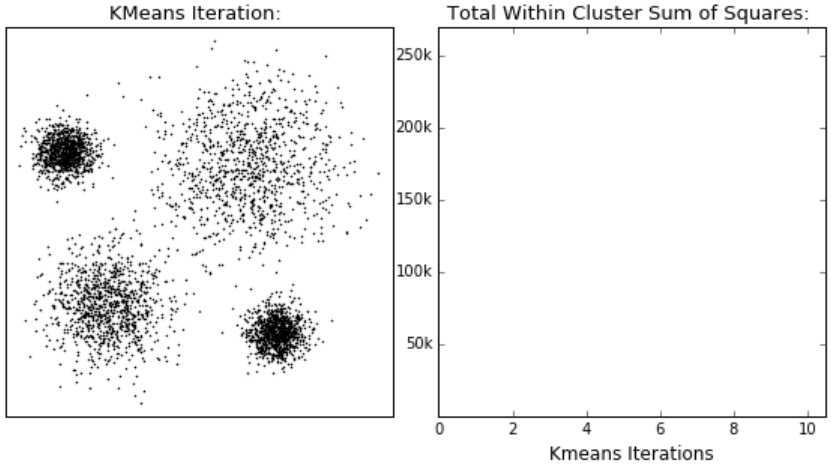
K=2

K=3

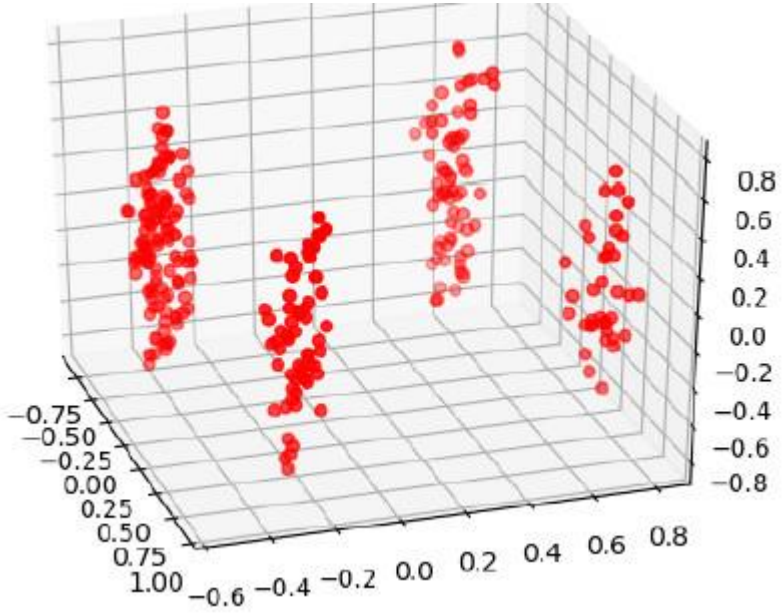


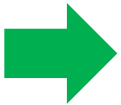
Elbow point

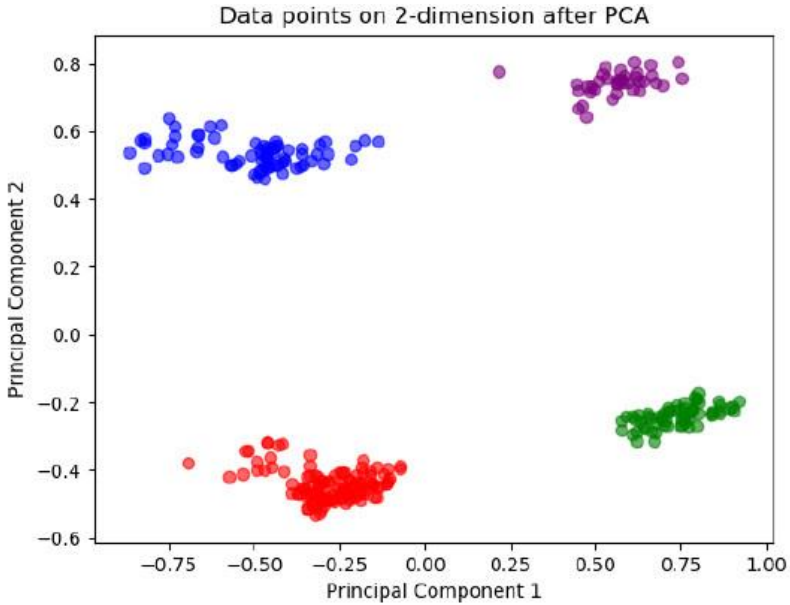




High Dimension Data

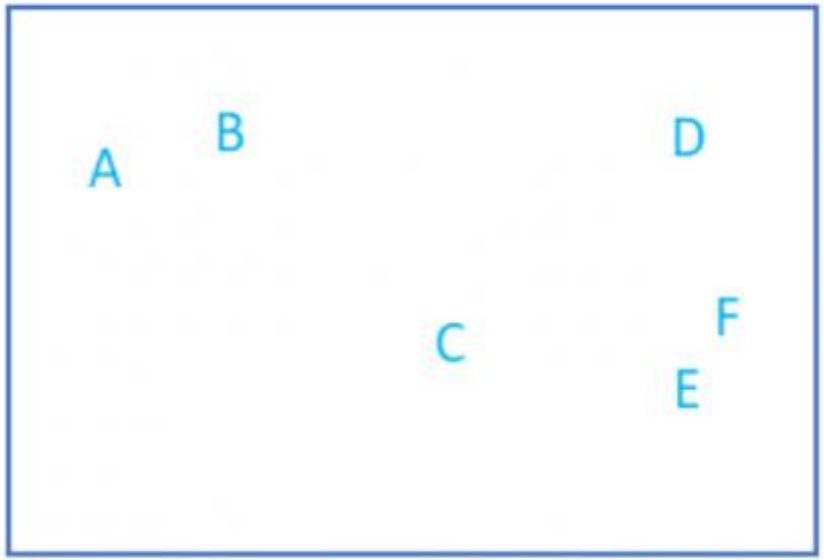




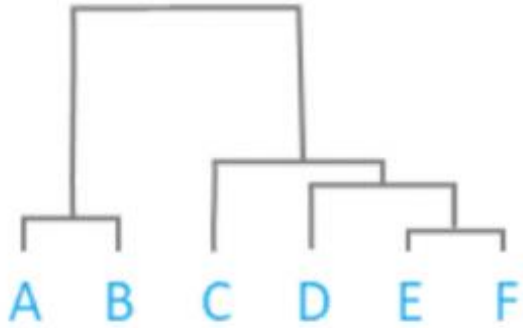


Hierarchical Methods

Garanti BBVA Technology

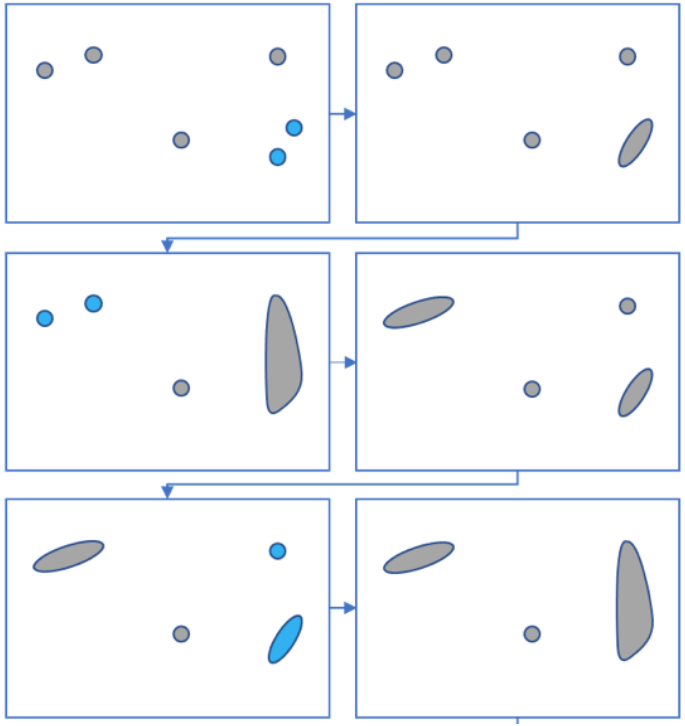


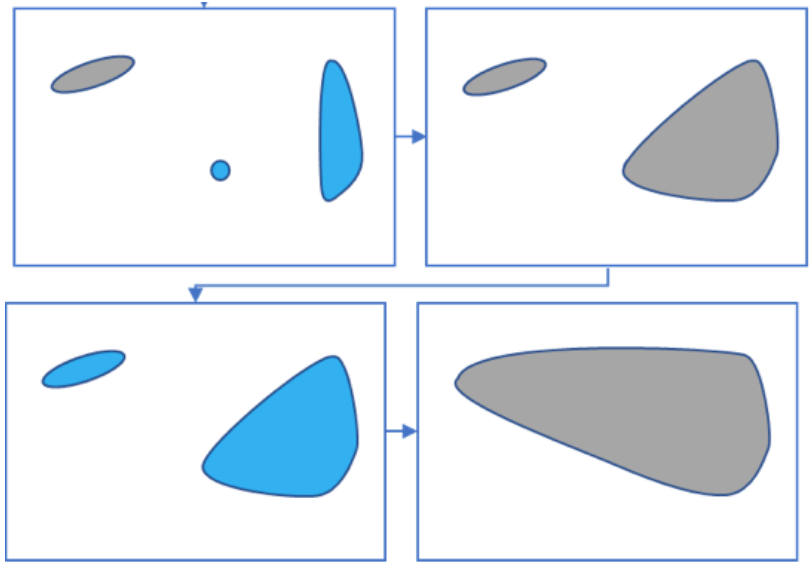
Dendrogram



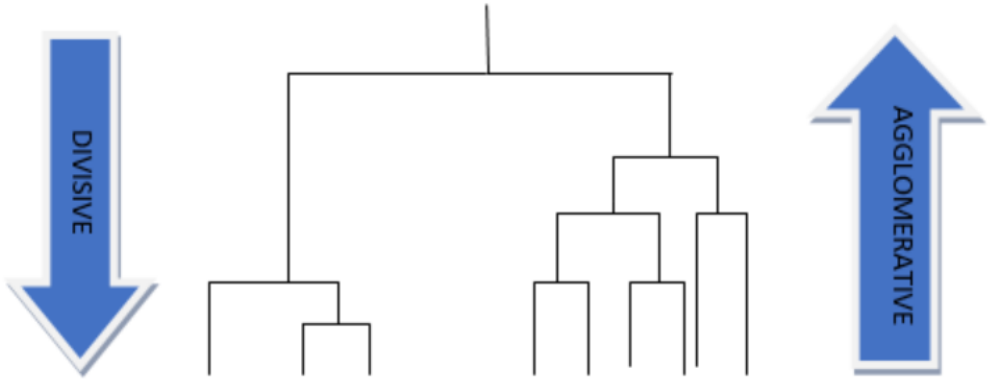
Identify the two clusters that are closest together

Merge the two most similar clusters





Hierarchical Methods

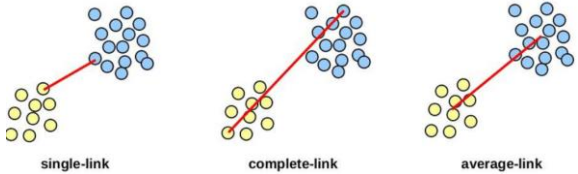


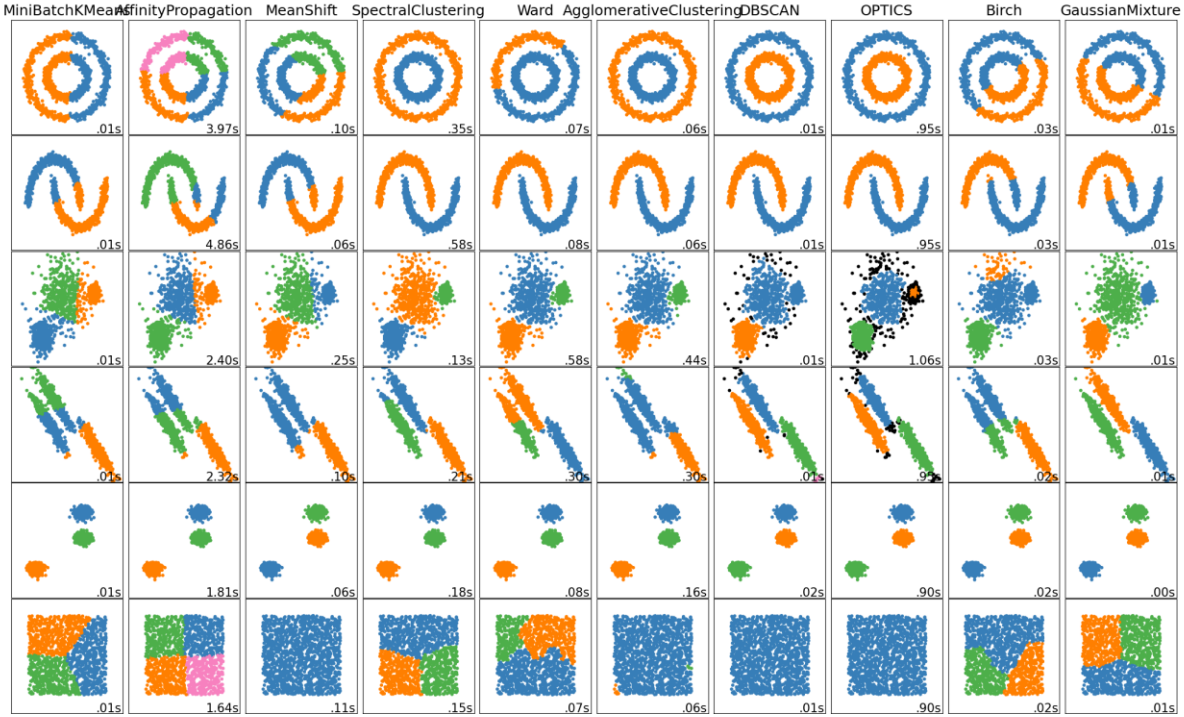
- Distance calculation between points
- Aggregate closest points
- In the beginning: n(sample size) cluster, in the end: 1 cluster if it is not stopped

Distance Measurements

Stopping Criteria

- The number of cluster
- Min distance between clusters





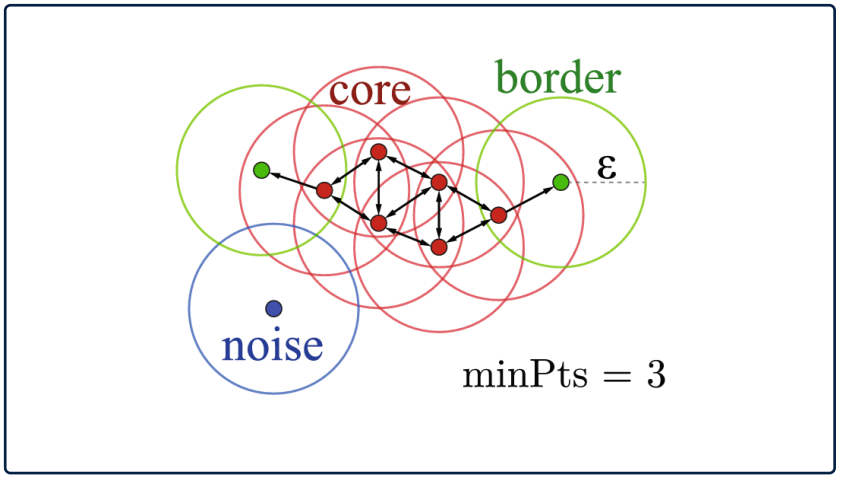
Density Based Methods

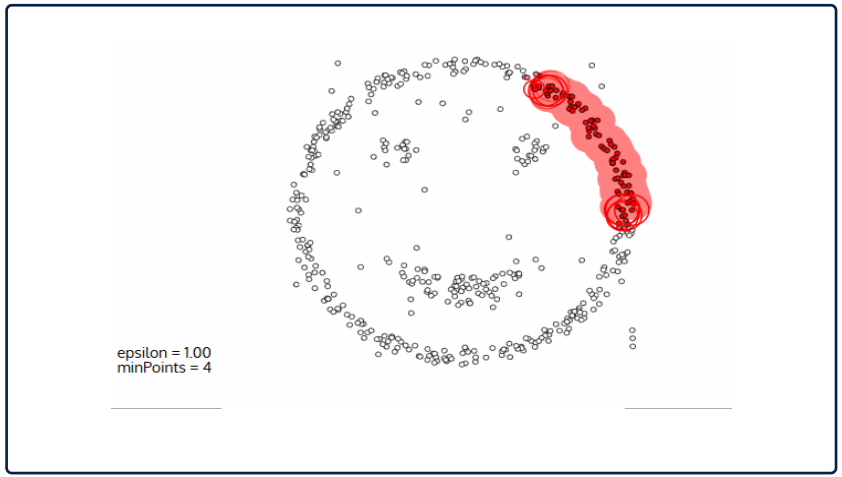
Garanti BBVA Technology

DBSCAN* Algorithm

Finds core samples of high density and expands clusters from them.

- **minPts:** The number of samples (or total weight) in a neighborhood for a point to be considered as a core point.
- **eps:** The maximum distance between two samples for one to be considered as in the neighborhood of the other.
- **Core**: This is a point that has at least m points within distance n from itself.
- **Border**: This is a point that has at least one Core point at a distance n.
- **Noise**: This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself.

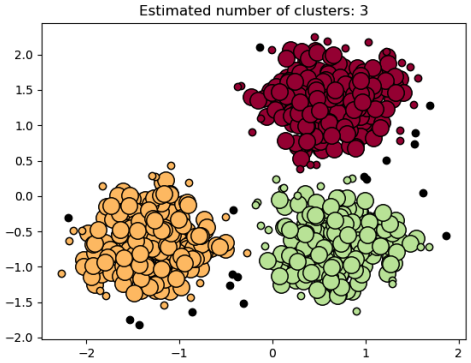


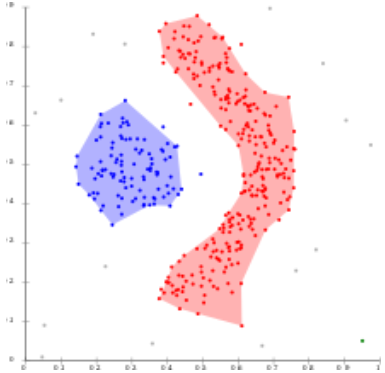


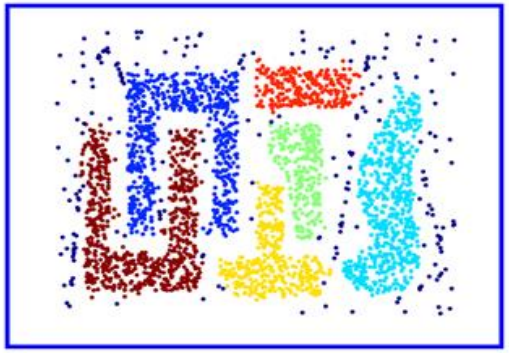
*__DBSCAN:__ **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise

DBSCAN - Output

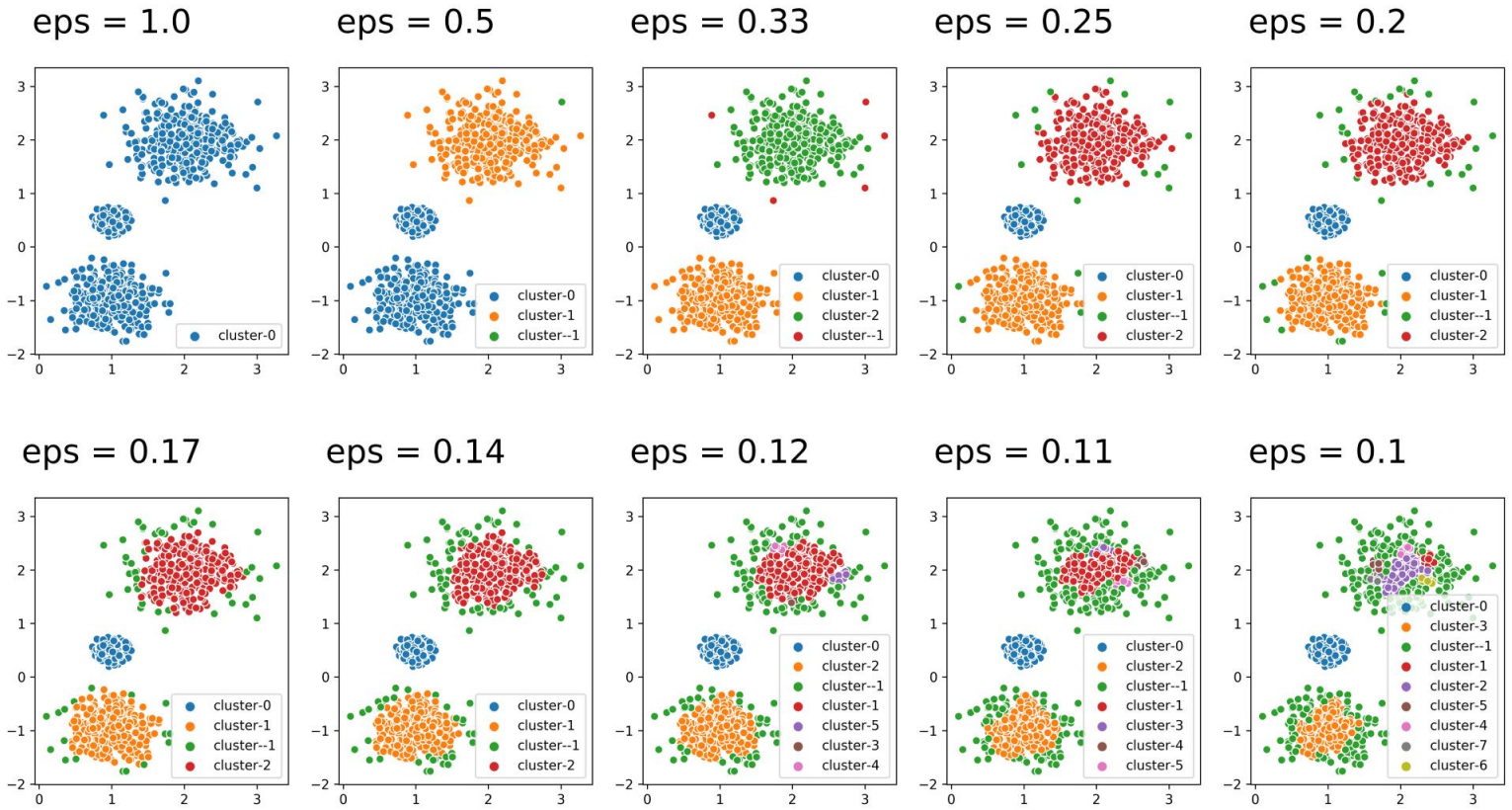
Can discover arbitrarily shaped clusters.



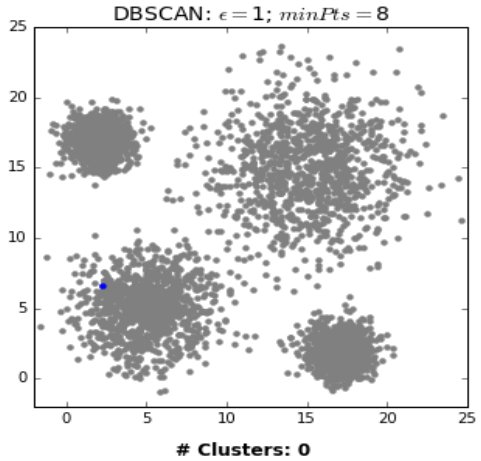


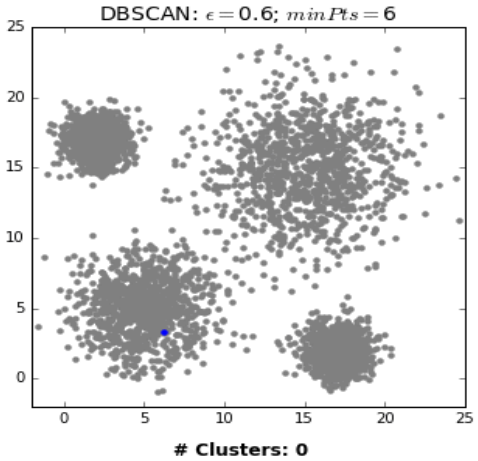


DBSCAN – Tuning Parameters

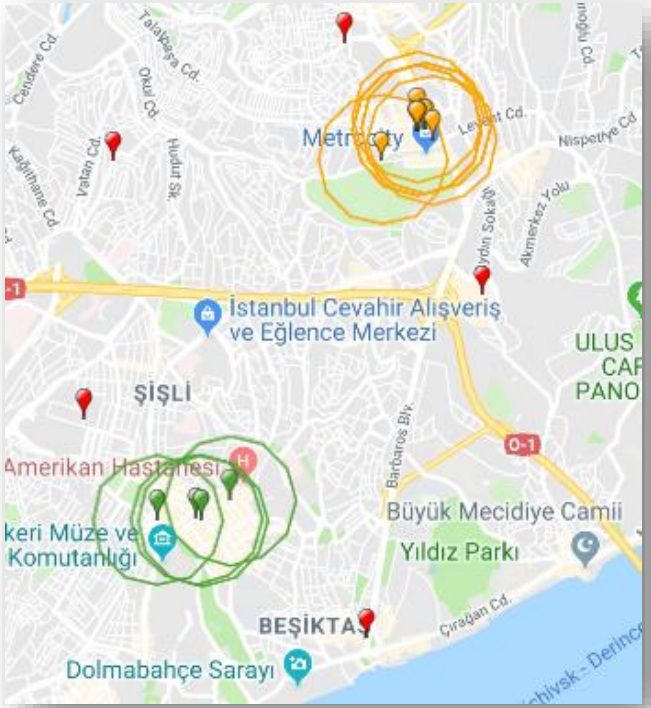


DBSCAN – Tuning Parameters





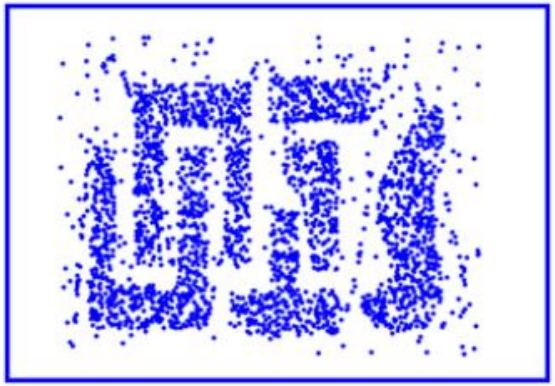
DBSCAN - Output

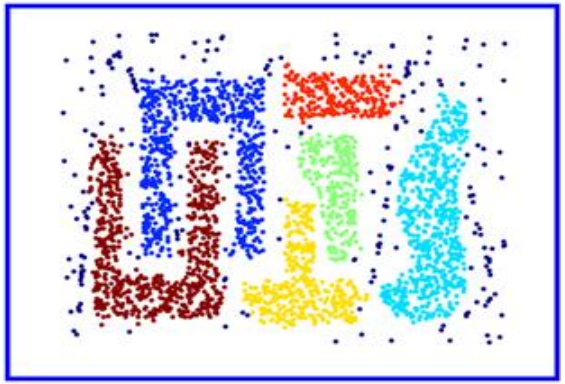


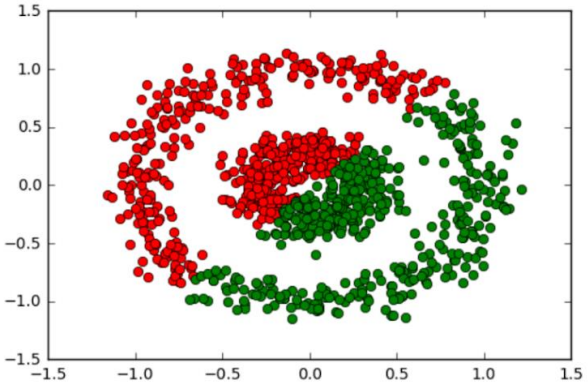
habitat habitat noise

habitat habitat habitat habitat noise

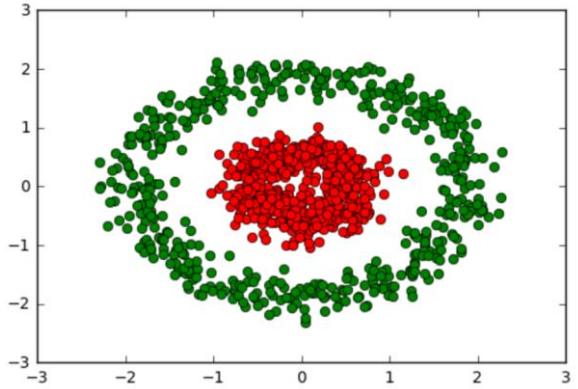
KMeans vs DBSCAN





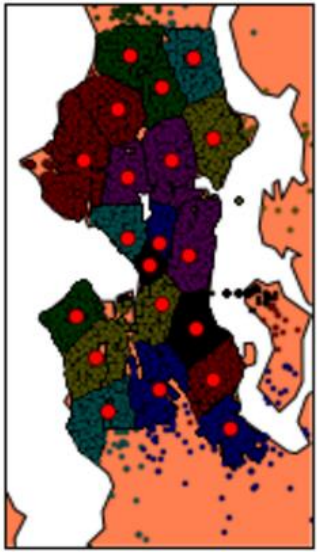


KMeans

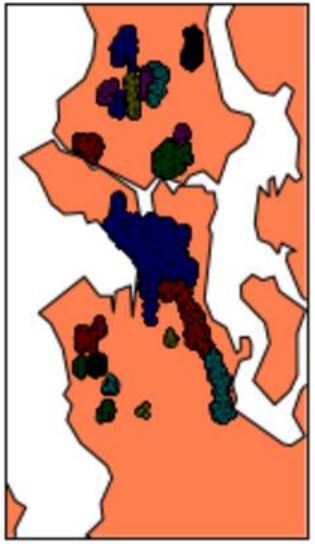


Dbscan

KMeans vs DBSCAN



KMeans



Dbscan

Applications, Advantages, Disadvantages of DBSCAN

Applications

- Location applications
- Recommendation systems
- Customer segmentation
- Outlier elimination

Advantages

- Does not require one to specify the number of clusters, as opposed to k-means.
- Can discover arbitrarily shaped clusters.
- Is great at separating clusters of high density versus clusters of low density within a given dataset.
- Is great with handling outliers within the dataset.
- The parameters minPts and ε can be set by a domain expert, if the data is well understood.

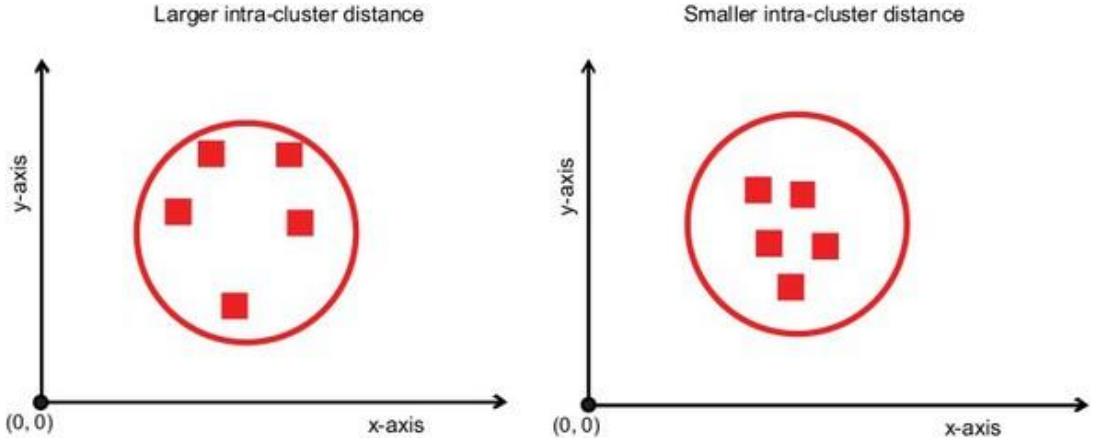
Disadvantages

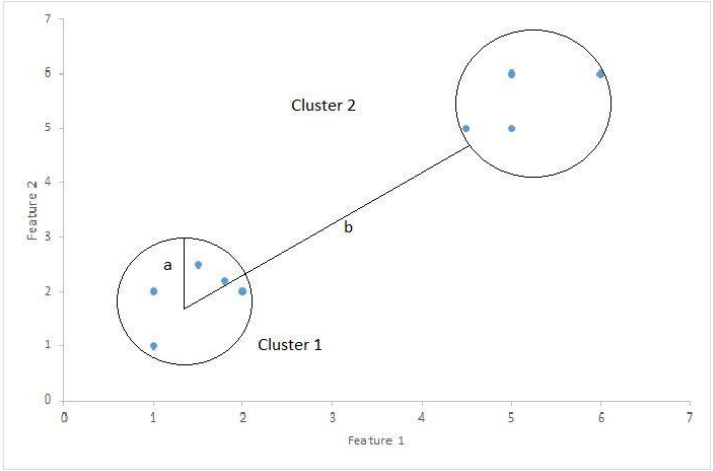
- Is not entirely deterministic(border points)
- The quality of DBSCAN depends on the distance measure.
- Struggles with high dimensionality data
- Computational complexity
- If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.

Interpreting Clustering Results

Preferred clusters

- Bigger distance between clusters
- Lower distance between data points in same clusters





Garanti BBVA Technology

Thanks