Mondragon Unibertsitatea

Escuela Politécnica Superior

Clustering

1

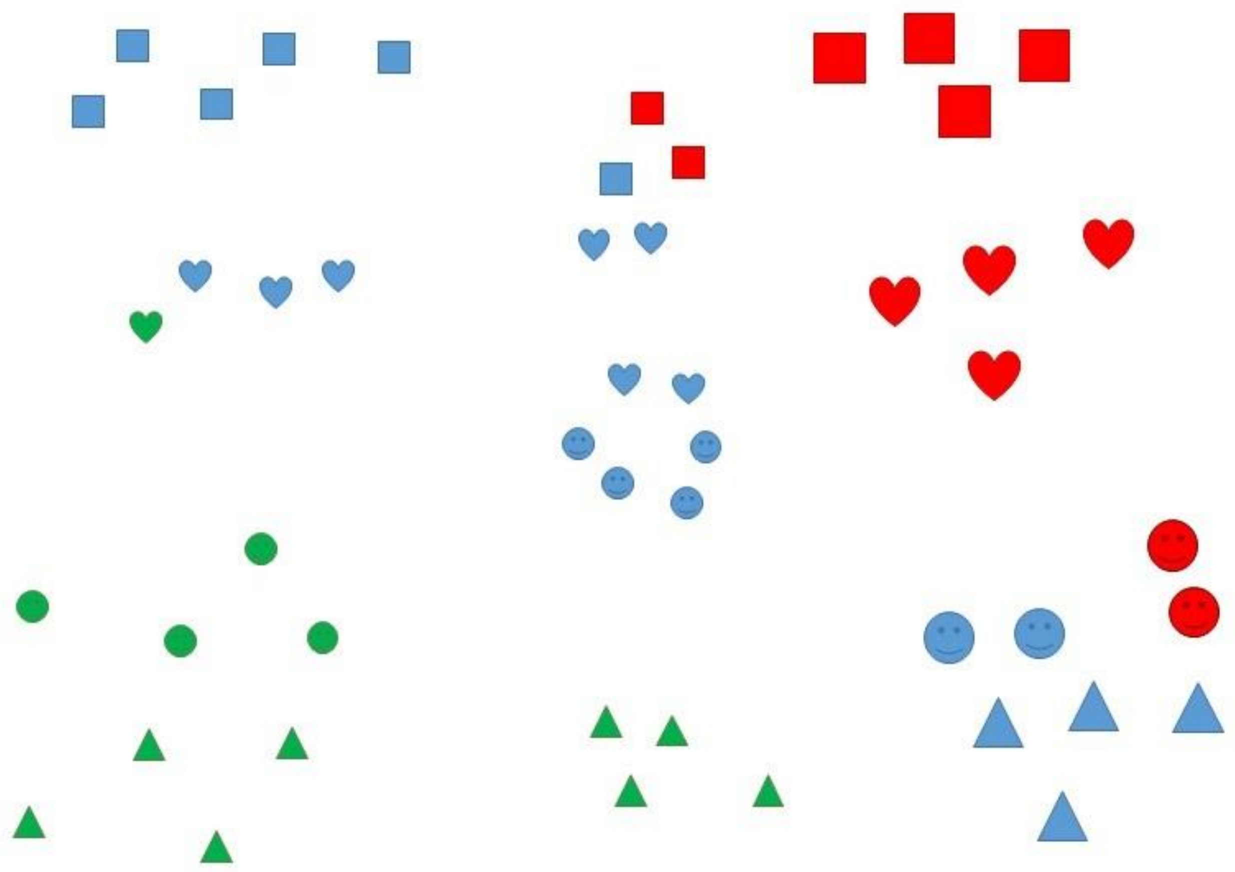
Introduction

Clustering

- Most frequent unsupervised method

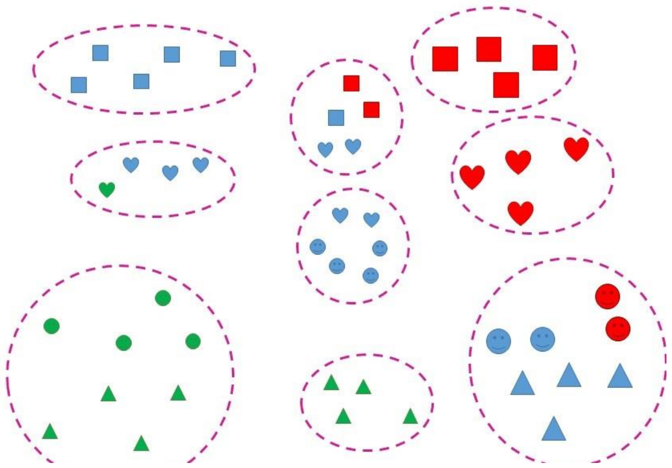
Clustering

Mondragon Unibertsitatea
Escuela Politécnica Superior

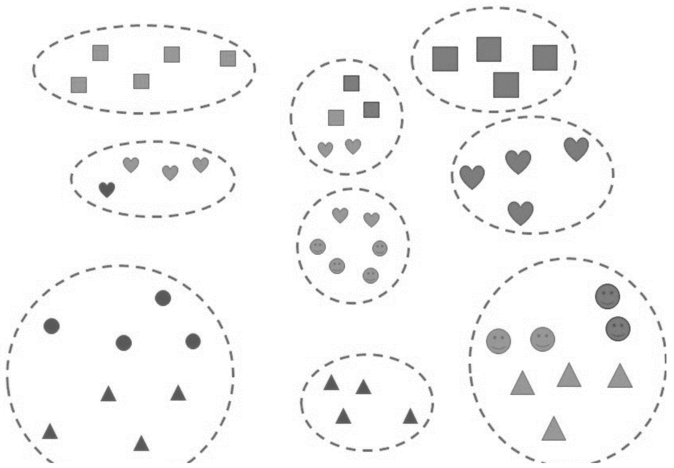


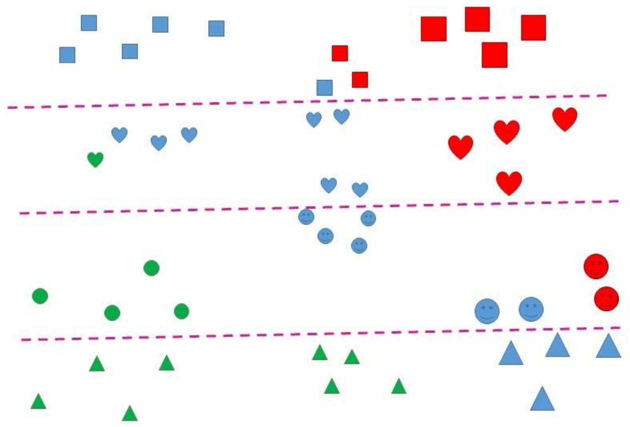
Clustering

Mondragon Unibertsitatea
Escuela Politécnica Superior

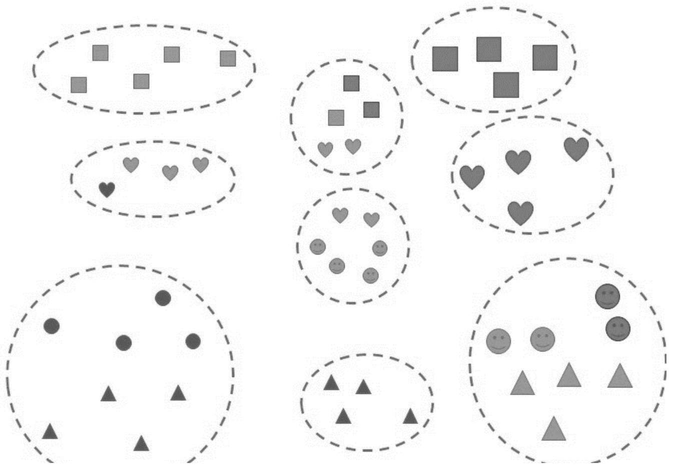


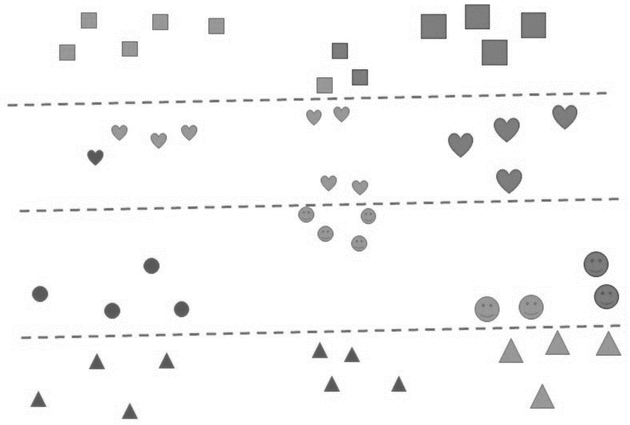
Clustering

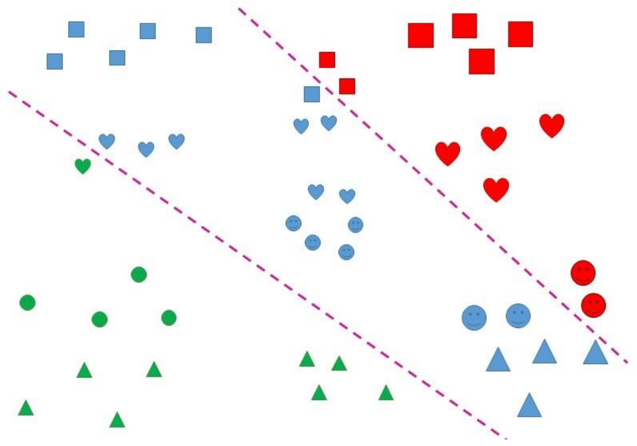




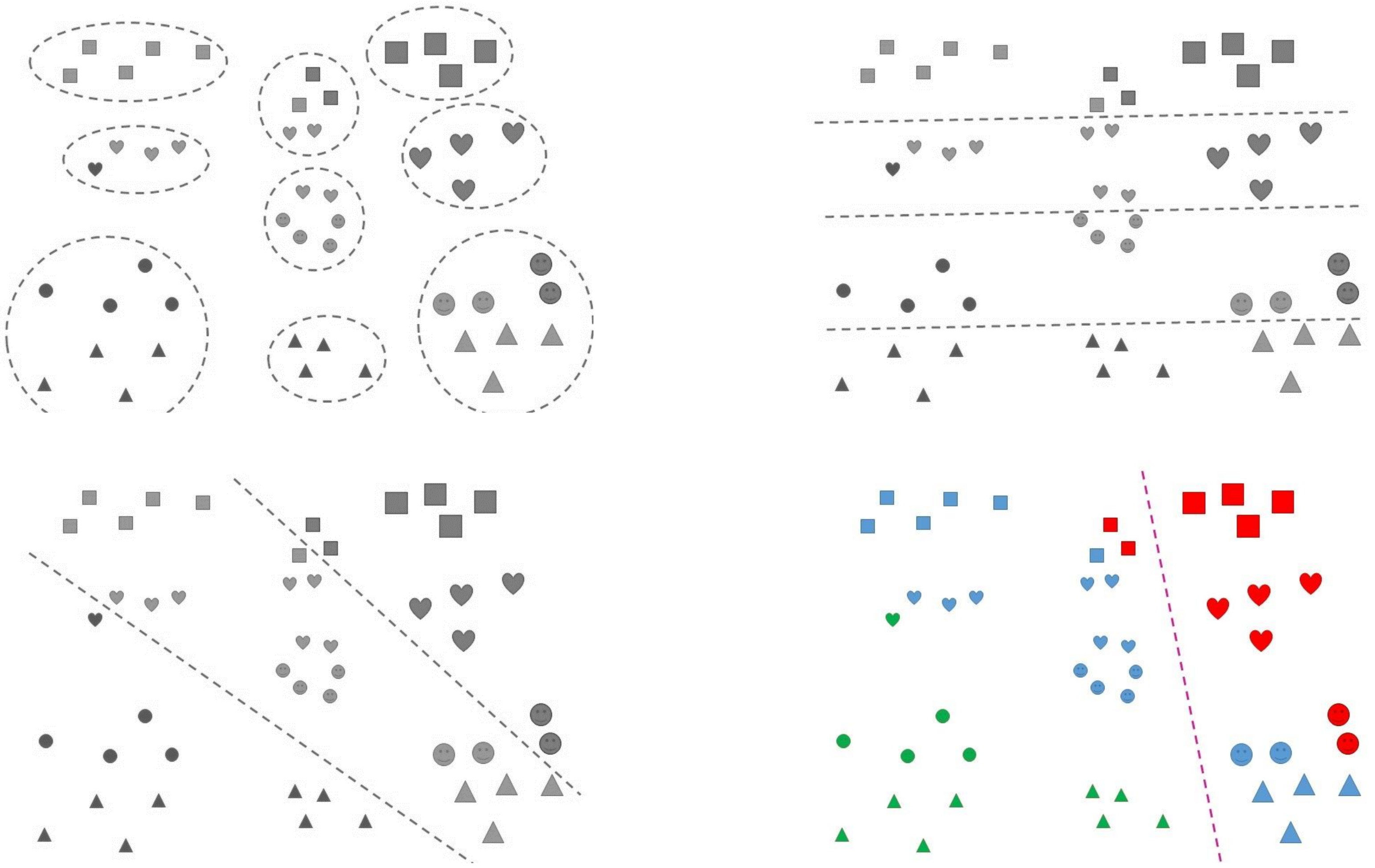
Clustering







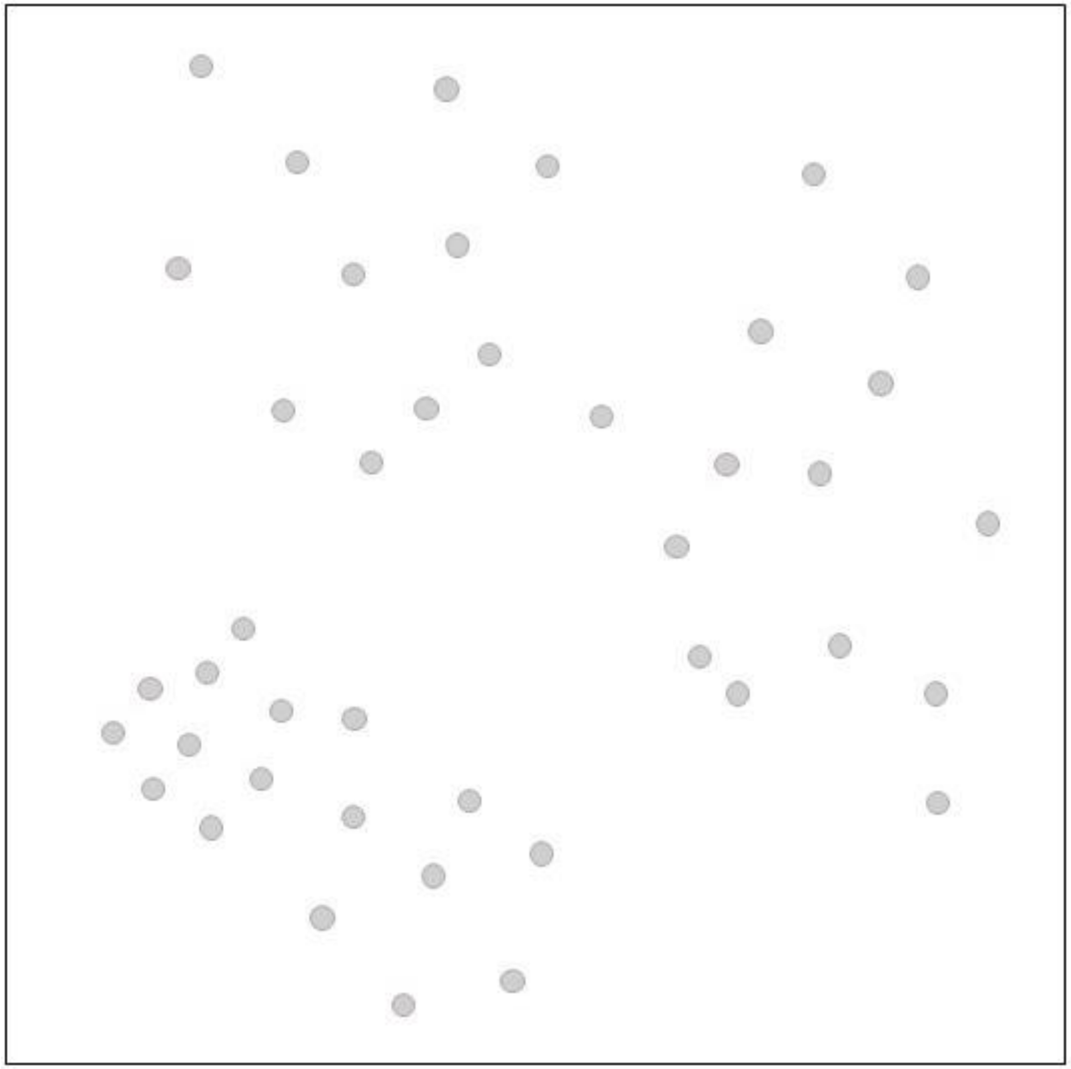
Clustering



2

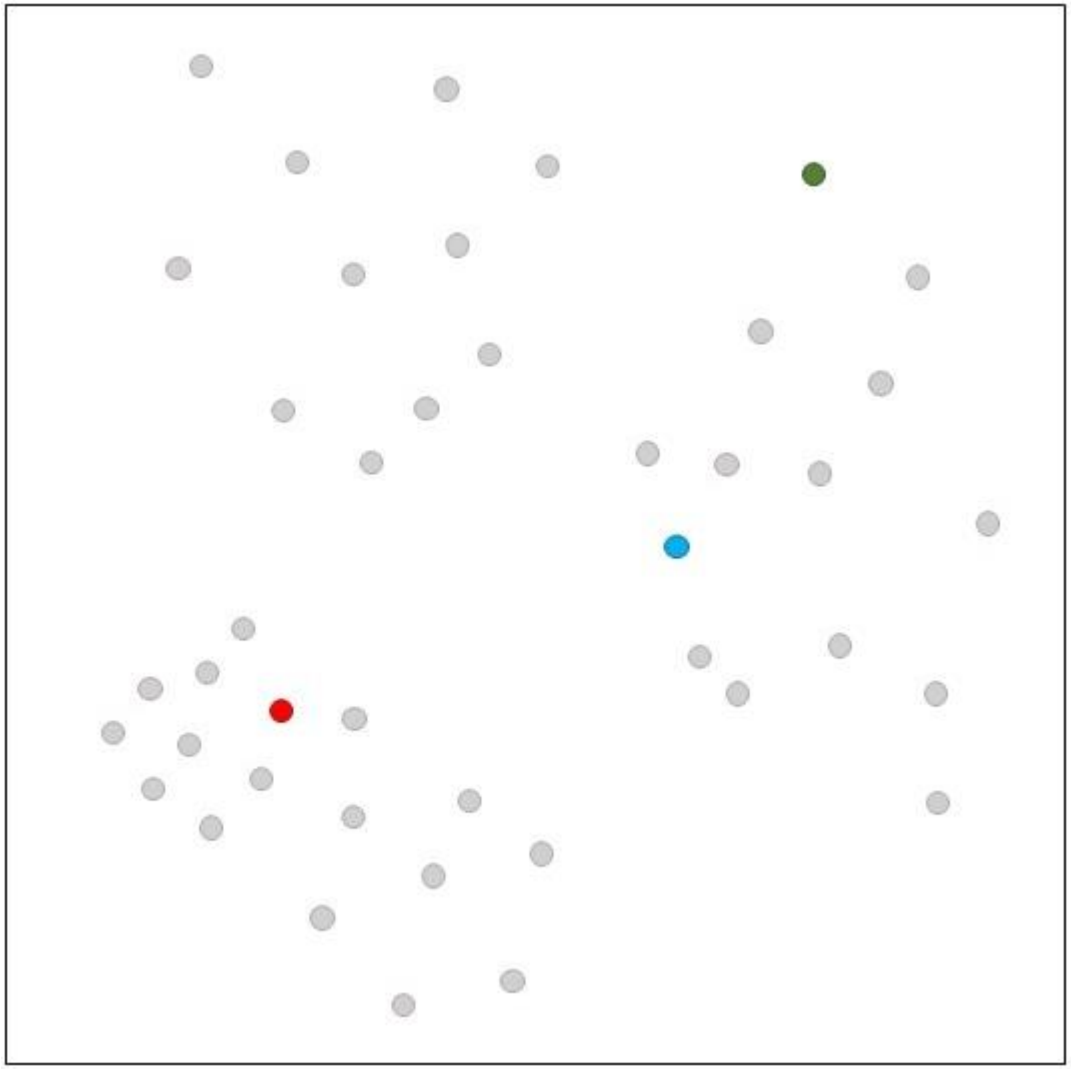
Algorithms

K-Means



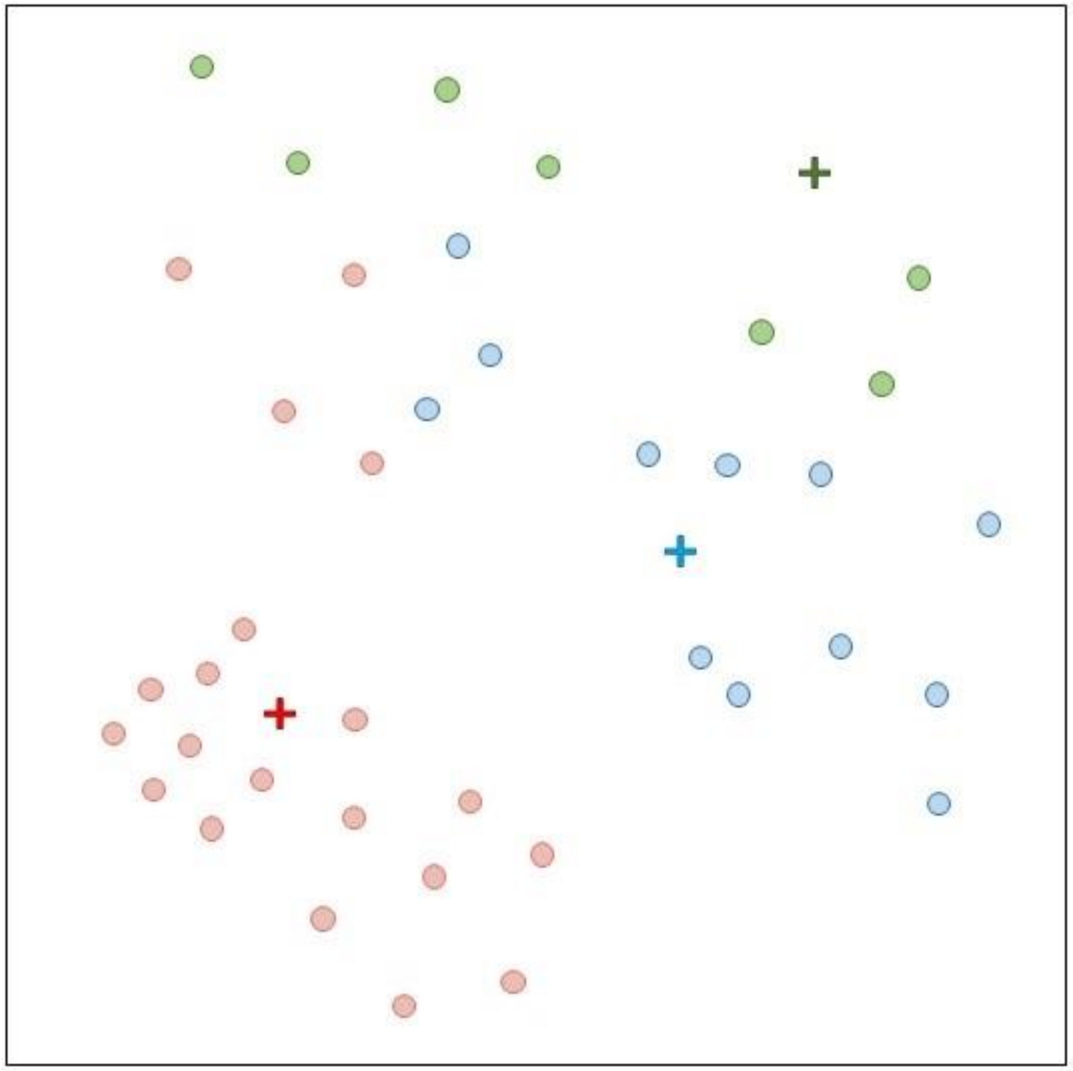
We prefix K and begin from the original data

K-Means



K points are randomly chosen (here $K = 3$) as initial centroids

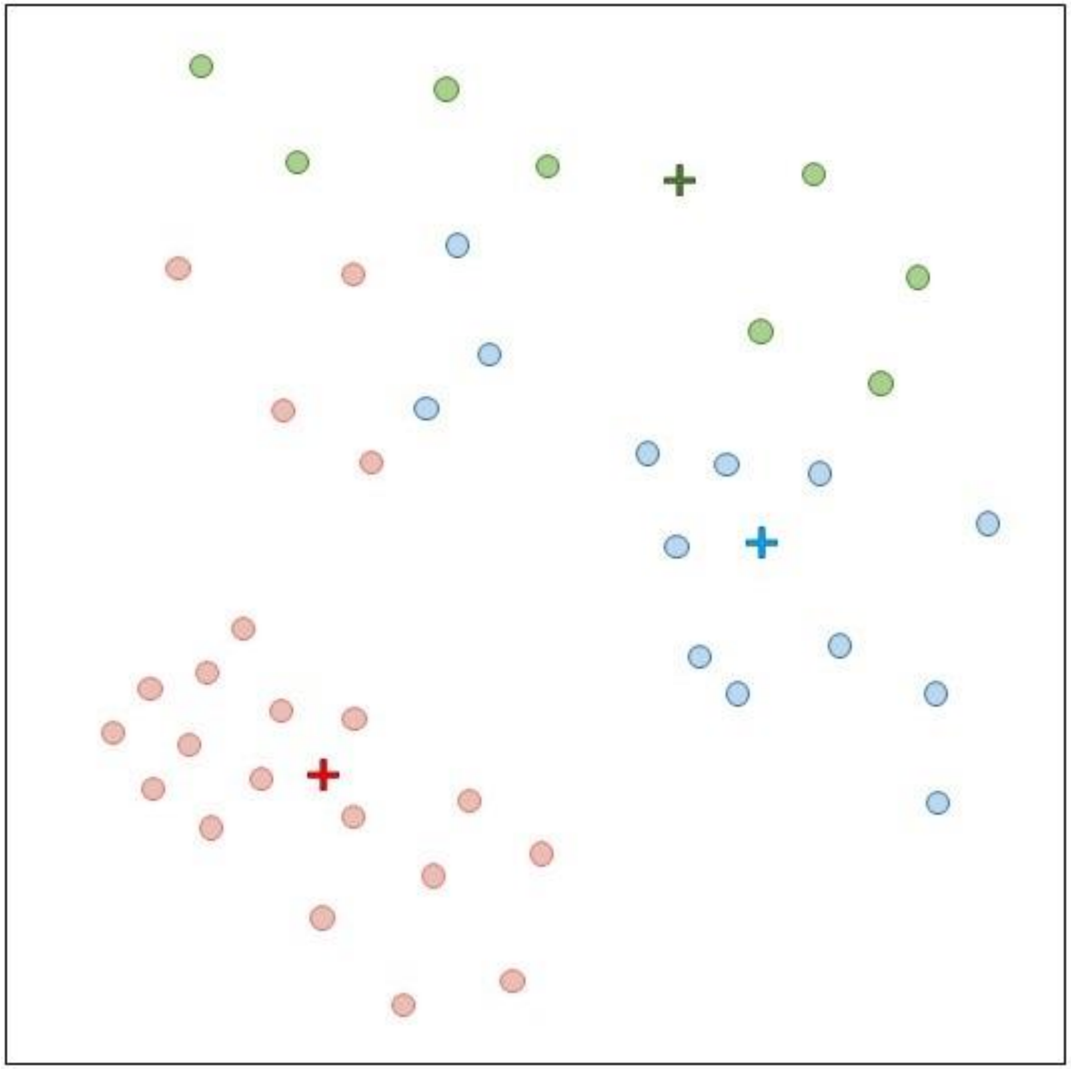
K-Means



a) All points are assigned to the cluster represented by the closest centroid

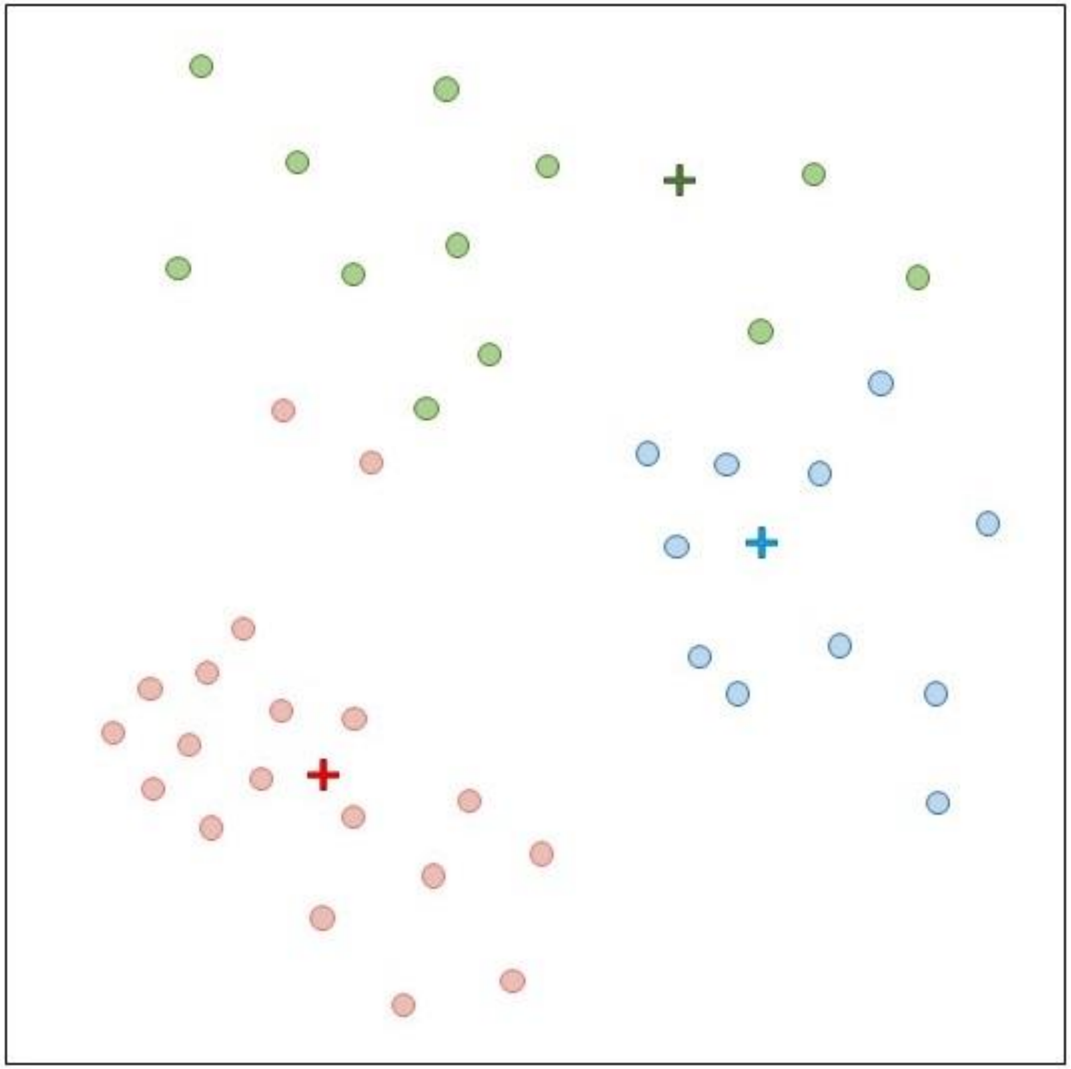
K-Means

Mondragon Unibertsitatea
Escuela Politécnica Superior



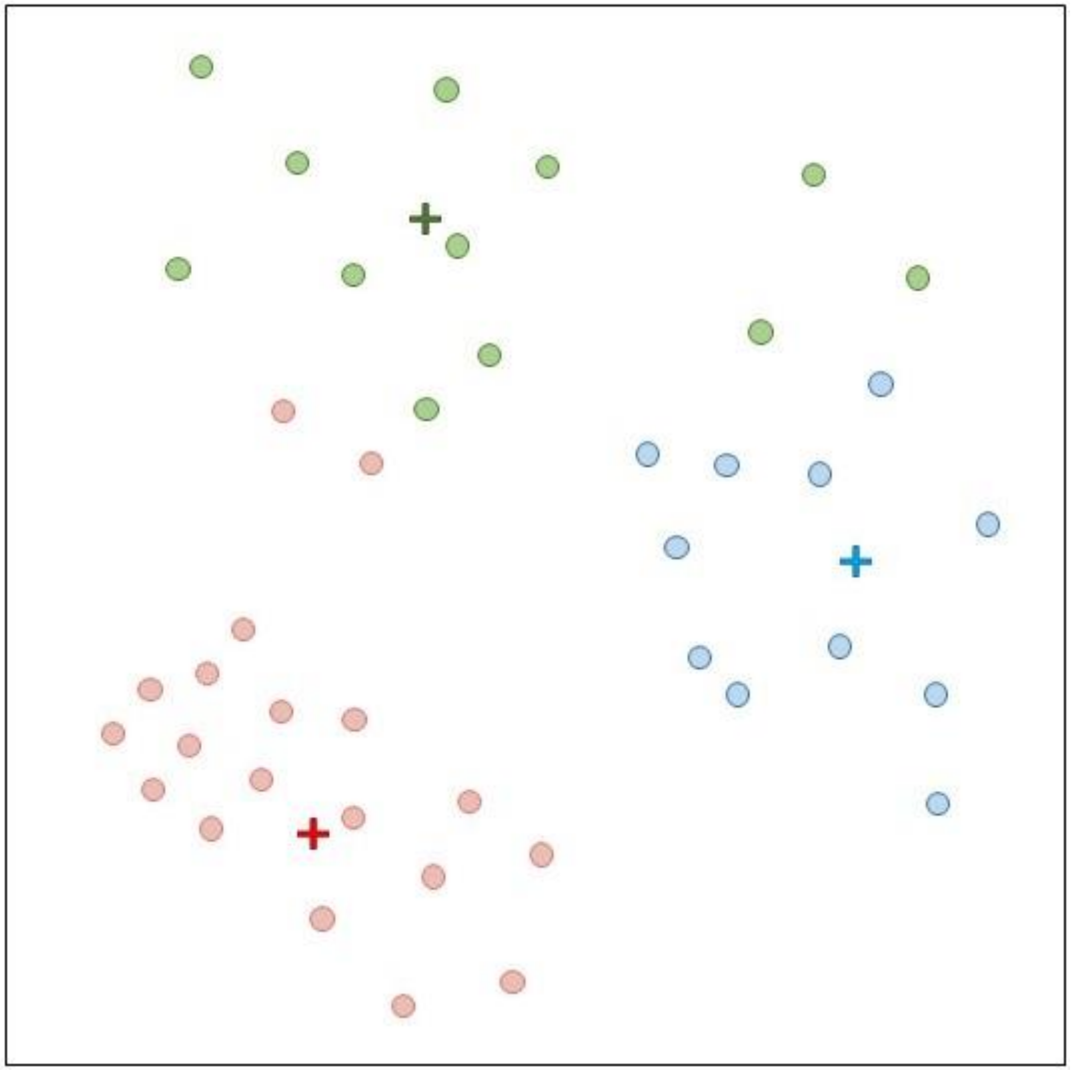
b) The centroid (mass center) of each cluster is obtained. It could be none of the points

K-Means

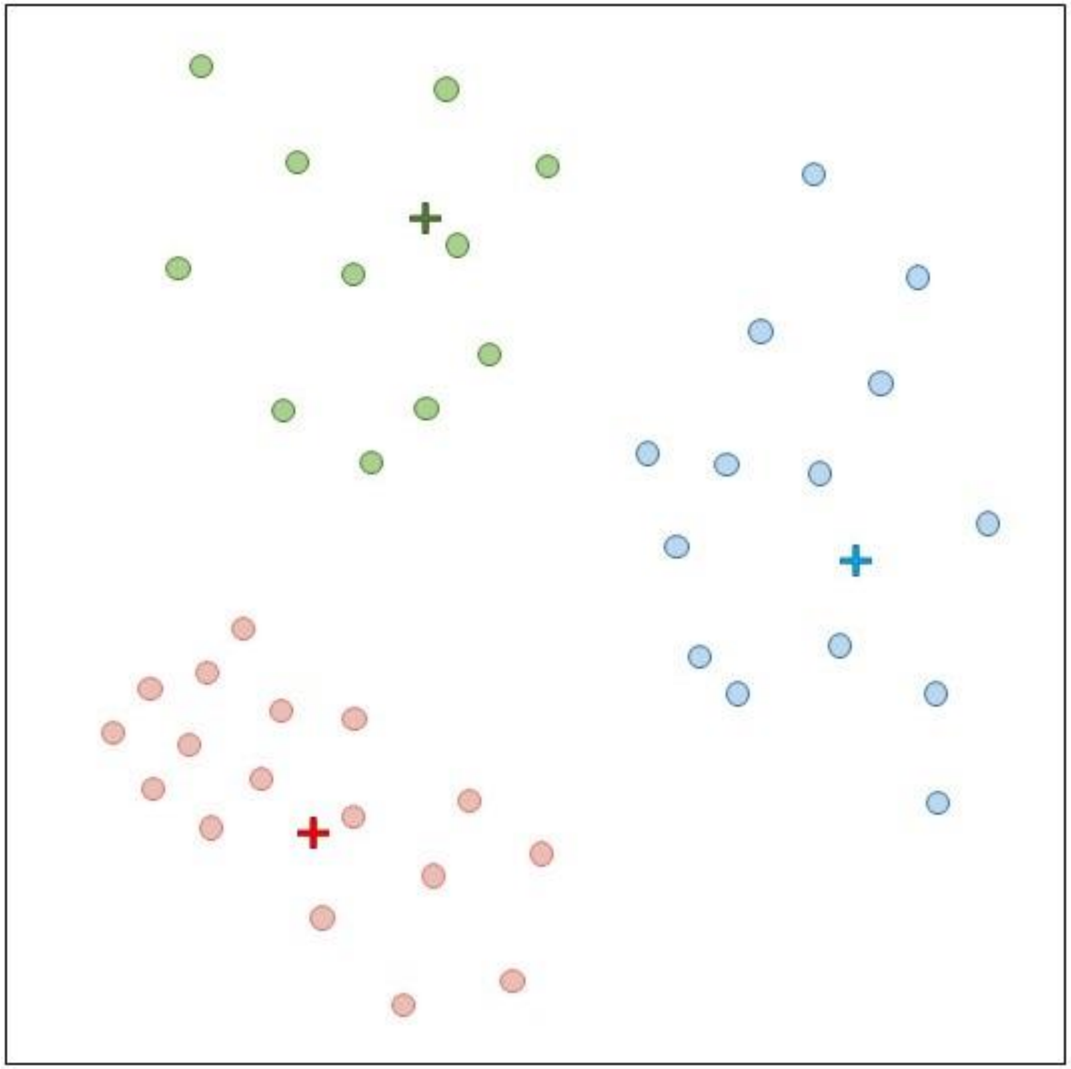


a) All points are assigned to the cluster represented by the closest centroid

K-Means

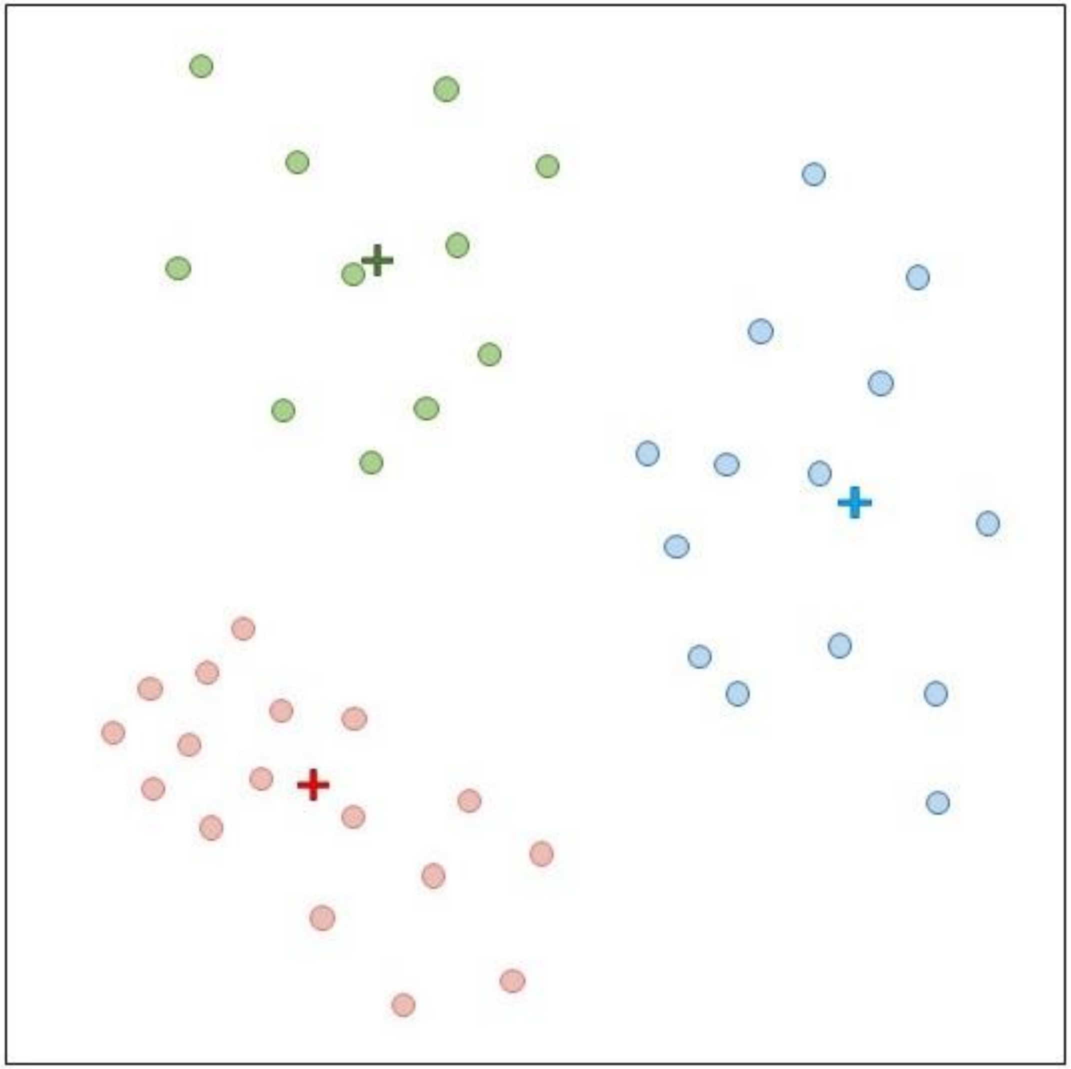

b) The centroid of each cluster is obtained

K-Means



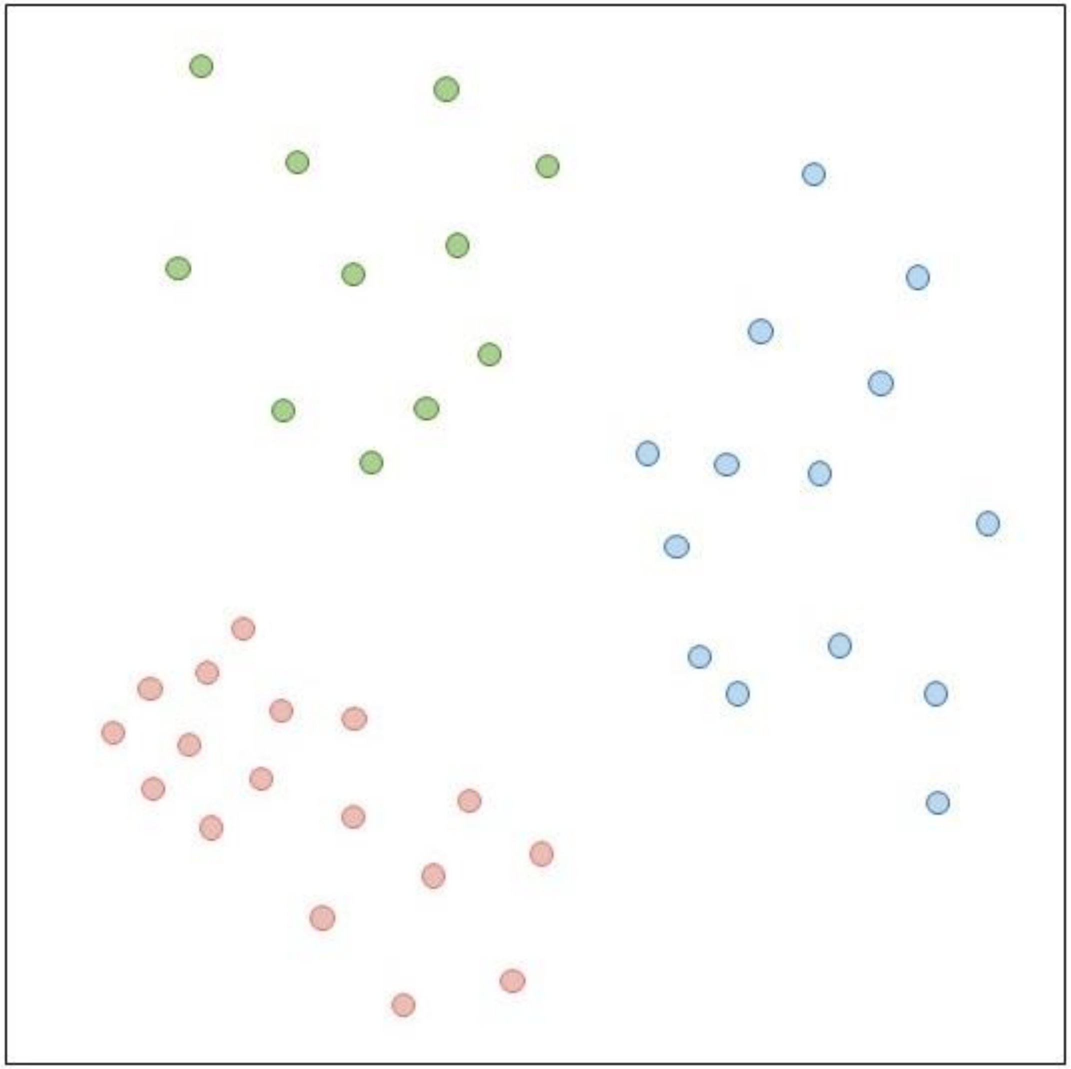
a) All points are assigned to the cluster represented by the closest centroid

K-Means



b) The centroid of each cluster is obtained

K-Means



We stop when no more changes in the centroids occur
(or if a prefixed limit number of iterations a-b is reached)

Affinity propagation

Mondragon Unibertsitatea

Escuela Politécnica Superior

- No need for prefixing the amount of clusters.

Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

- No need for prefixing the amount of clusters.
- It combines concepts

Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.

Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.

Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k

Affinity propagation

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and

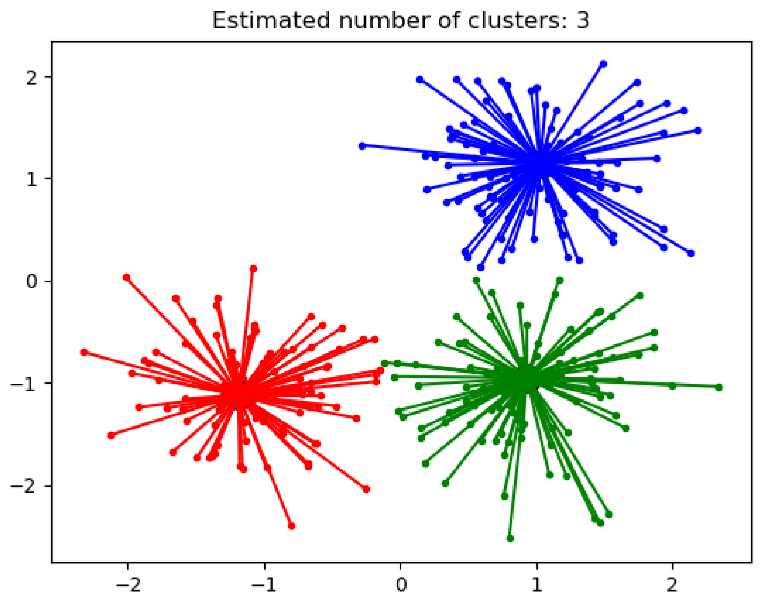
Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and
  - (2) has been chosen for many samples as example to follow.

Affinity propagation

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and
  - (2) has been chosen for many samples as example to follow.

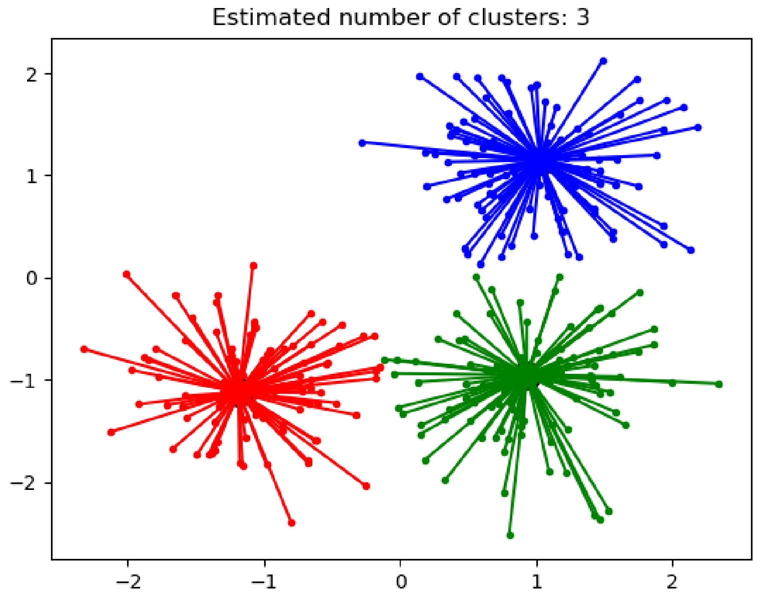


Affinity propagation

Mondragon Unibertsitatea
Escuela Politécnica Superior

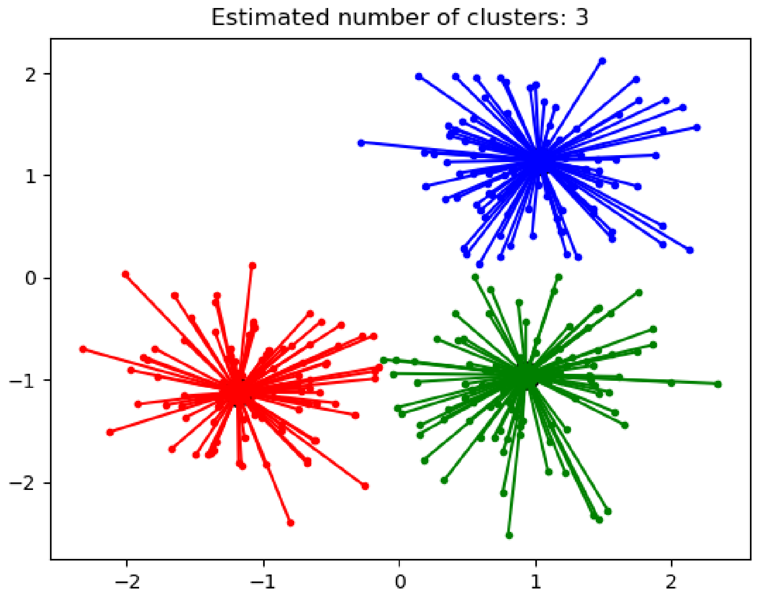
- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and
  - (2) has been chosen for many samples as example to follow.

- It is an iterative process in which "messages" are sent, catching affinities and recalculating preferences until convergence or reaching a predefined iterations limit.



Affinity propagation

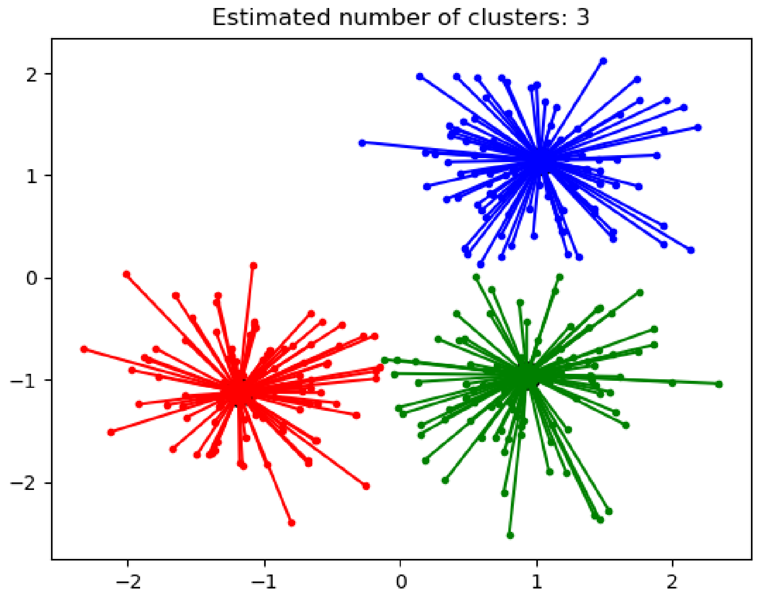
- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and
  - (2) has been chosen for many samples as example to follow.



- It is an iterative process in which “messages” are sent, catching affinities and recalculating preferences until convergence or reaching a predefined iterations limit.
- It is a very complex method.

Affinity propagation

- No need for prefixing the amount of clusters.
- It combines concepts:
  - **Responsability**: Cummulative evidence of sample k should be an example for sample i. Potential examples and their relationships are studied.
  - **Availability**: Cummulative evidence of sample i should choose sample k as an example. Sample i is compared with the rest of possible examples separately.
- Then, sample i choose to follow sample k if k
  - (1) is similar enough (affine) to many samples, and
  - (2) has been chosen for many samples as example to follow.



- It is an iterative process in which “messages” are sent, catching affinities and recalculating preferences until convergence or reaching a predefined iterations limit.
- It is a very complex method.
- Its usage is recommended to be restricted to small or medium datasets.

Mean shift

Mondragon Unibertsitatea
Escuela Politécnica Superior

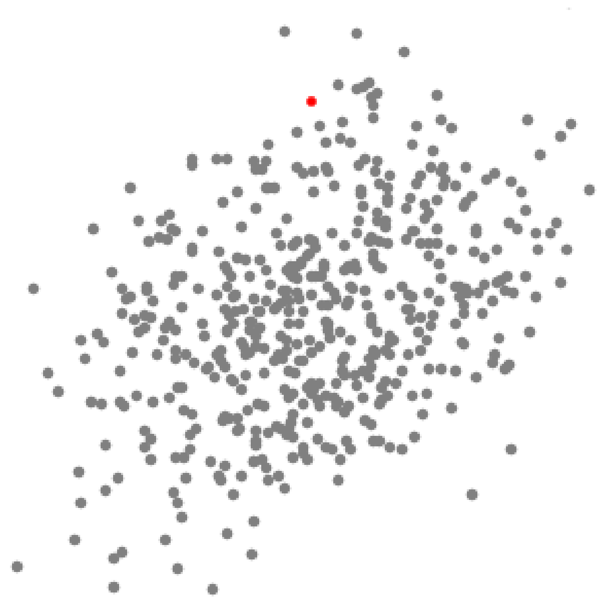
- No need for prefixing the amount of clusters.

Mean shift

- No need for prefixing the amount of clusters.
- It tries to search for “blobs“ (manchas) in high density areas by iteratively moving the centers

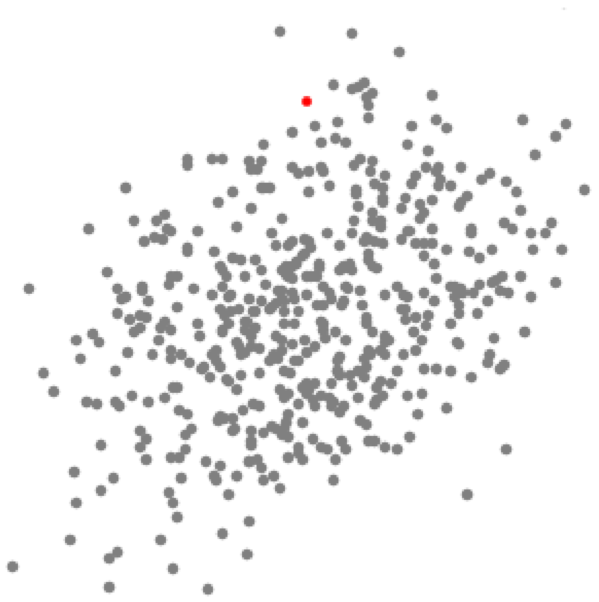
Mean shift

- No need for prefixing the amount of clusters.
- It tries to search for “blobs“ (manchas) in high density areas by iteratively moving the centers
- Computacionally expensive, as plenty of distance calculations are needed



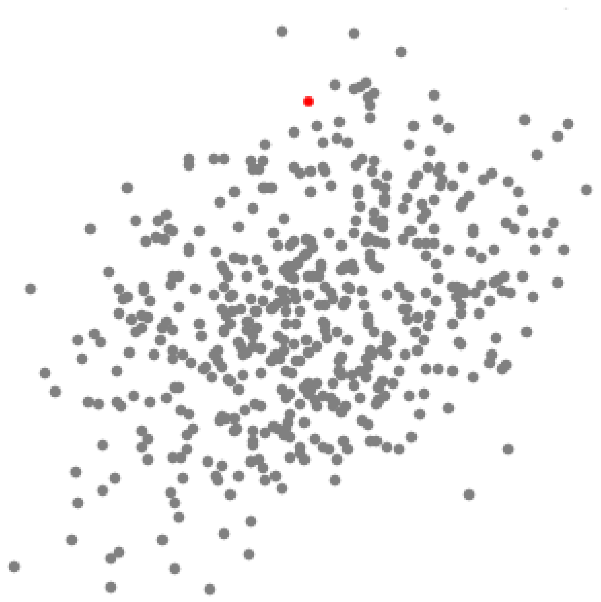
Mean shift

- No need for prefixing the amount of clusters.
- It tries to search for “blobs“ (manchas) in high density areas by iteratively moving the centers
- Computacionally expensive, as plenty of distance calculations are needed
- Only recommended for small or medium dense datasets



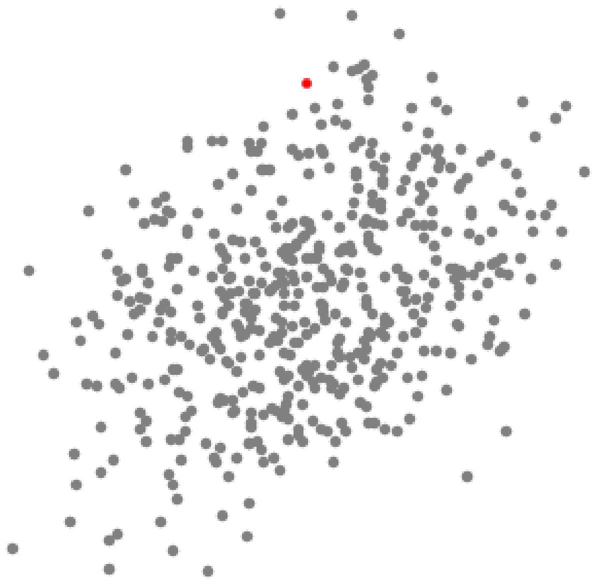
Mean shift

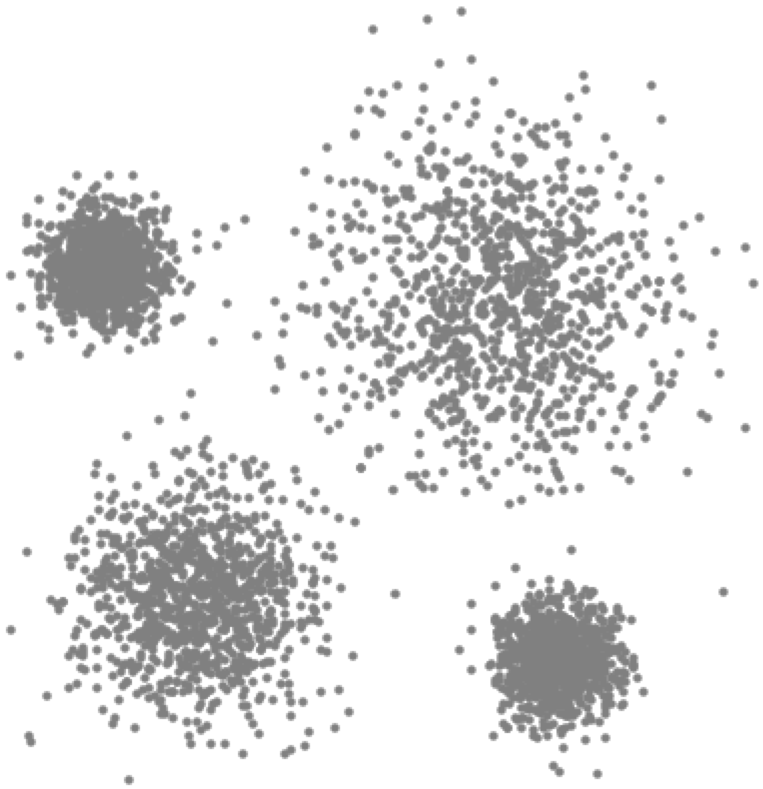
- No need for prefixing the amount of clusters.
- It tries to search for “blobs“ (manchas) in high density areas by iteratively moving the centers
- Computacionally expensive, as plenty of distance calculations are needed
- Only recommended for small or medium dense datasets
- A parameter called *bandwidth* limits the searching areas, thus influencing the amount of clusters. It can be prefixed or estimated during training



Mean shift

- No need for prefixing the amount of clusters.
- It tries to search for “blobs“ (manchas) in high density areas by iteratively moving the centers
- Computacionally expensive, as plenty of distance calculations are needed
- Only recommended for small or medium dense datasets
- A parameter called *bandwidth* limits the searching areas, thus influencing the amount of clusters. It can be prefixed or estimated during training





Hierarchical clustering

- No need for prefixing the amount of clusters

Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways

Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points

Hierarchical clustering

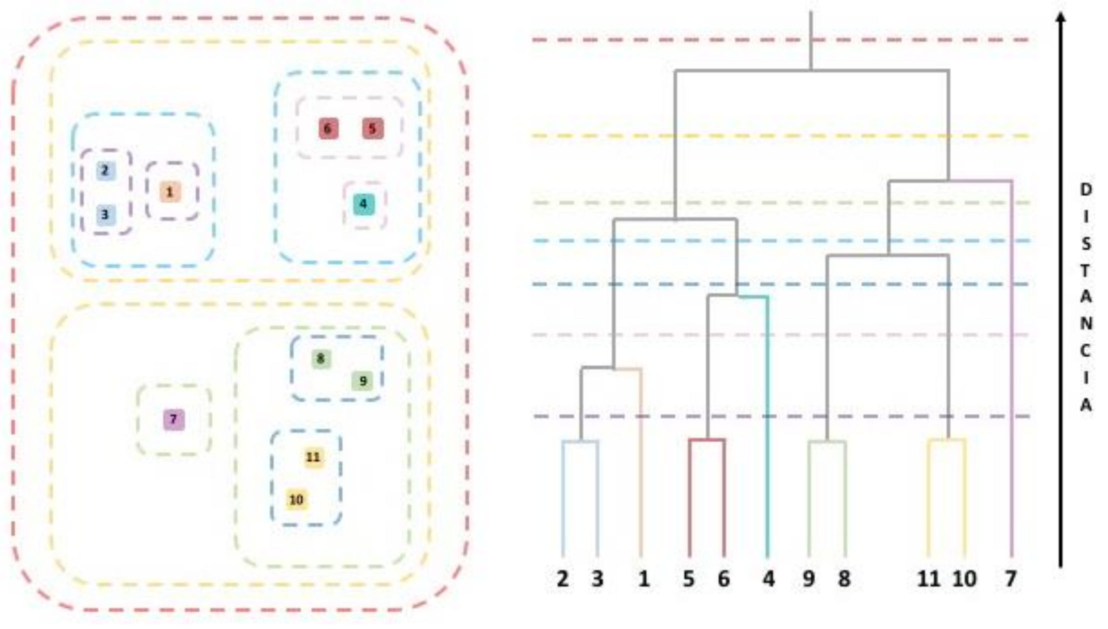
- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points
  - Top to bottom: There is an initial big cluster containing all points. Sequential splits are performed so that the resulting clusters are as much separated as possible

Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points
  - Top to bottom: There is an initial big cluster containing all points. Sequential splits are performed so that the resulting clusters are as much separated as possible
- In both cases, a summary graph (called *dendogram*) provides all the information about the successive clusters and their distances/similarities

Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points
  - Top to bottom: There is an initial big cluster containing all points. Sequential splits are performed so that the resulting clusters are as much separated as possible
- In both cases, a summary graph (called *dendogram*) provides all the information about the successive clusters and their distances/similarities

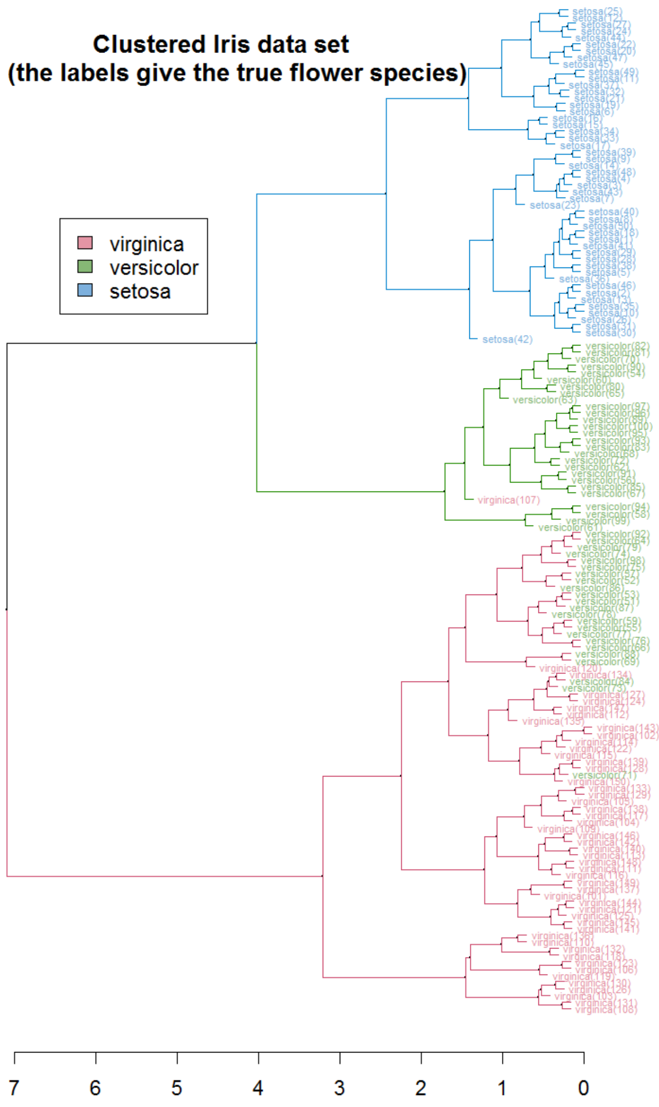


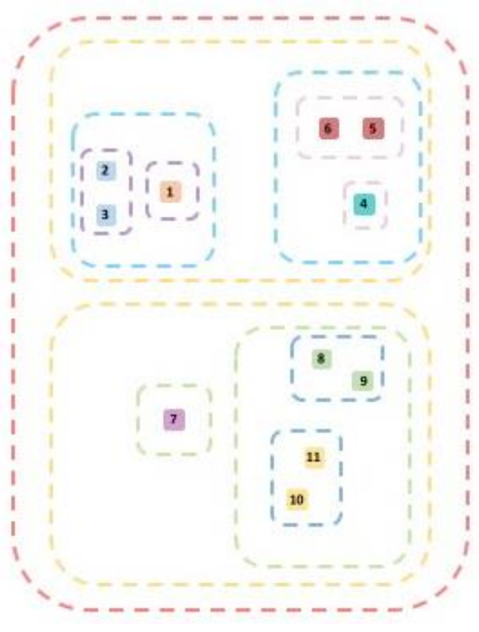
Mondragon Unibertsitatea

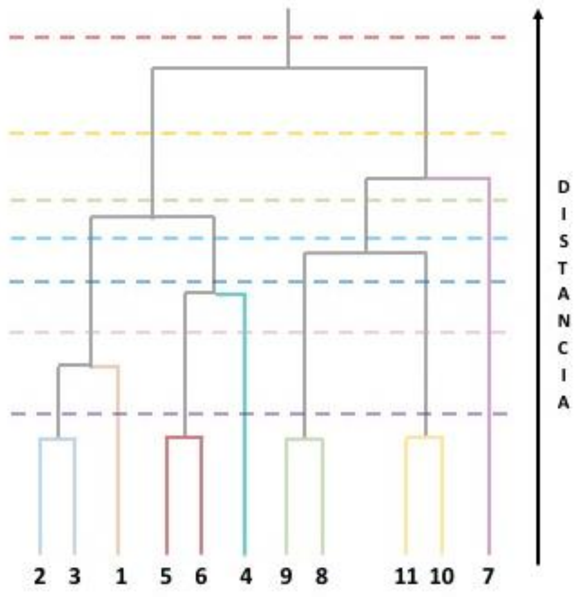
Escuela Politécnica Superior

Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points
  - Top to bottom: There is an initial big cluster containing all points. Sequential splits are performed so that the resulting clusters are as much separated as possible
- In both cases, a summary graph (called *dendogram*) provides all the information about the successive clusters and their distances/similarities

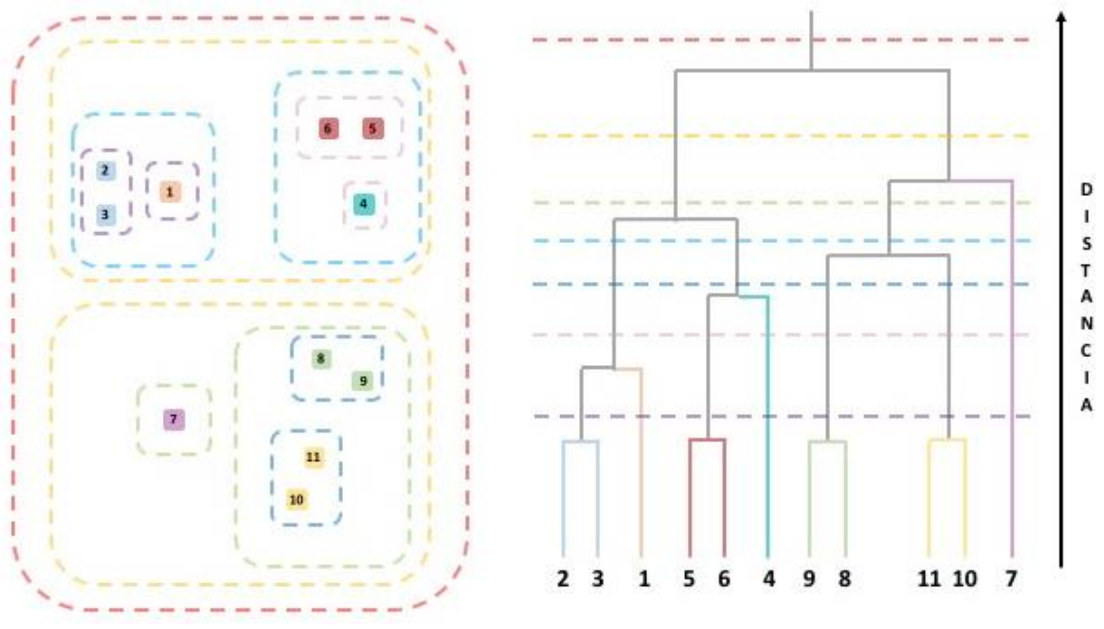


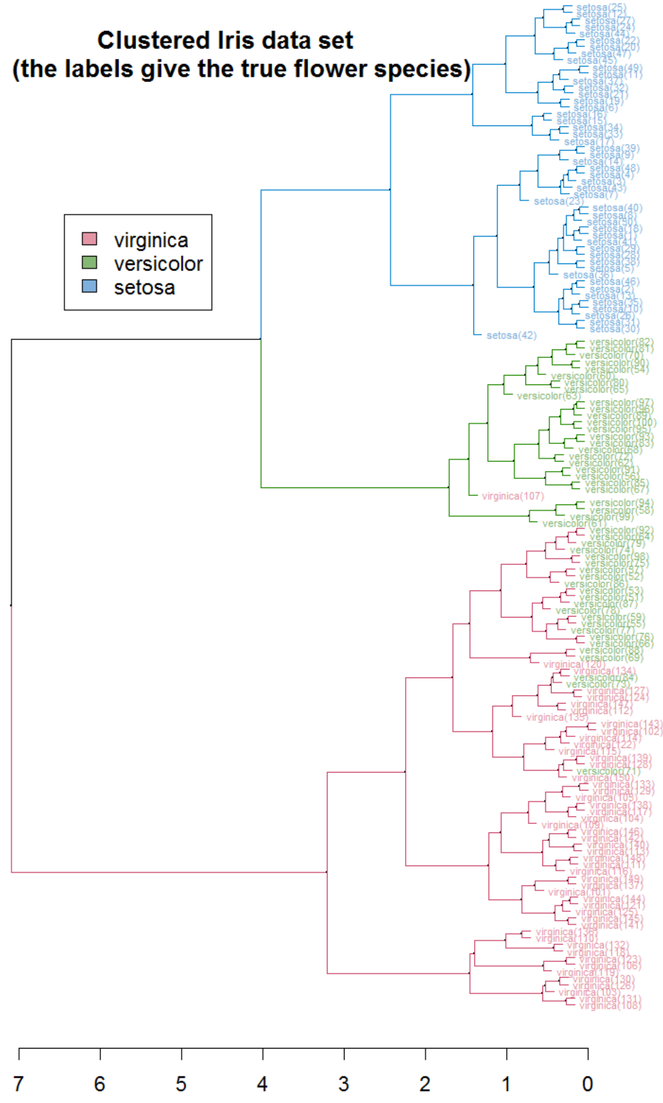




Hierarchical clustering

- No need for prefixing the amount of clusters
- The procedure flows in one of these two ways:
  - Bottom to top: Each single point begins being a cluster. Close clusters are sequentially merged until there is one single cluster containing all points
  - Top to bottom: There is an initial big cluster containing all points. Sequential splits are performed so that the resulting clusters are as much separated as possible
- In both cases, a summary graph (called *dendogram*) provides all the information about the successive clusters and their distances/similarities
- In order to be able to perform an early-stopping operation in the process, top to bottom approach is usually preferred





DBSCAN

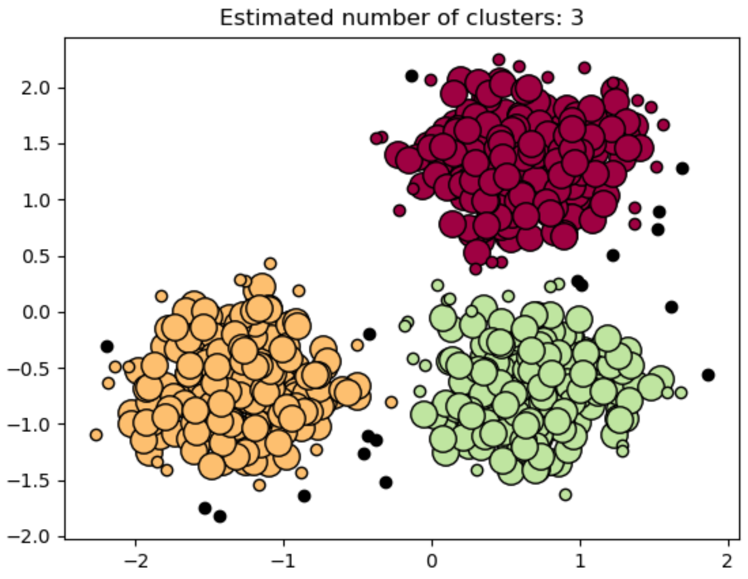
- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.

DBSCAN

- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.

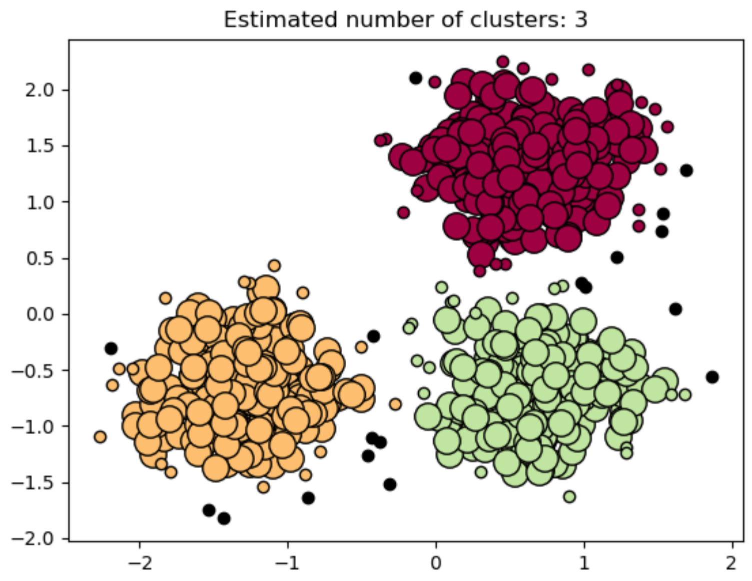
DBSCAN

- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.



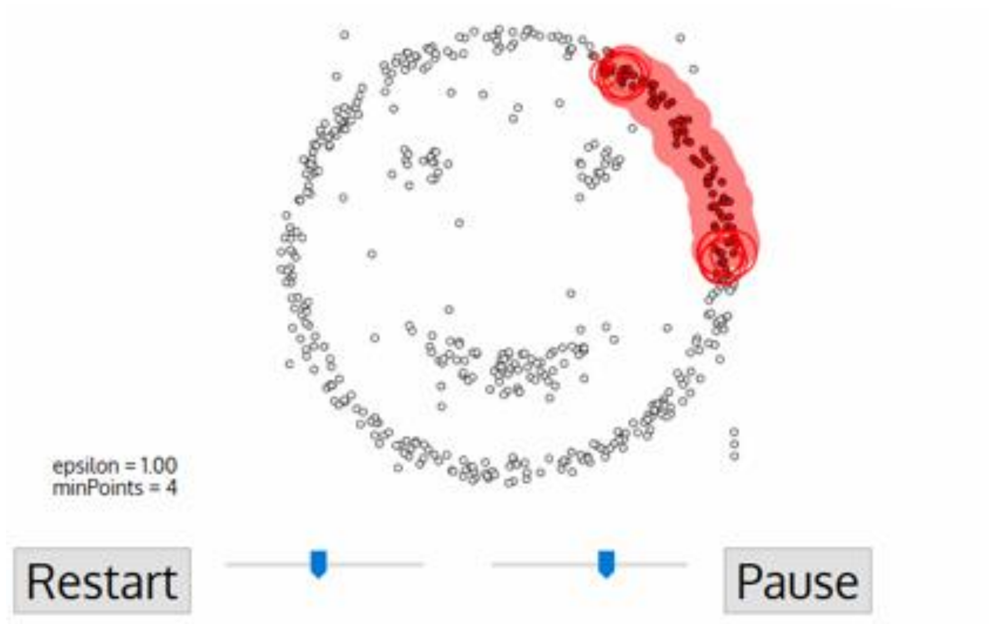
DBSCAN

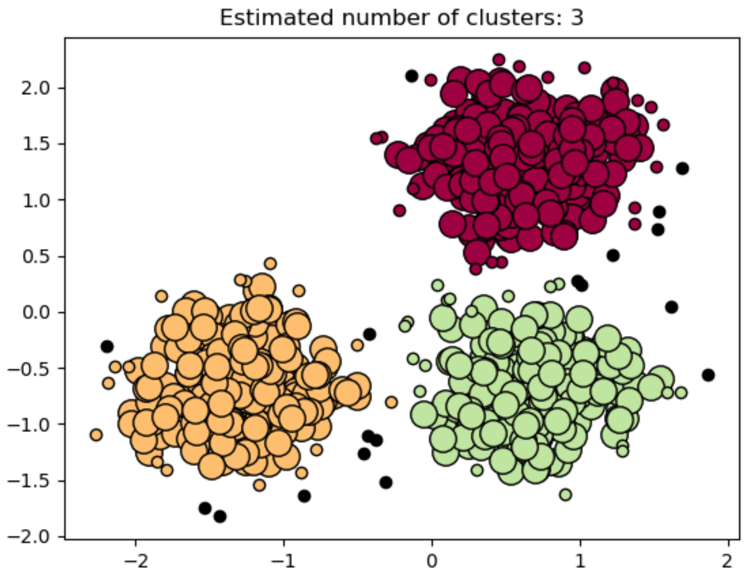
- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.
- A big difference with K-Means is that DBSCAN can define **clusters of arbitrary shape**, not only convex sets.



DBSCAN

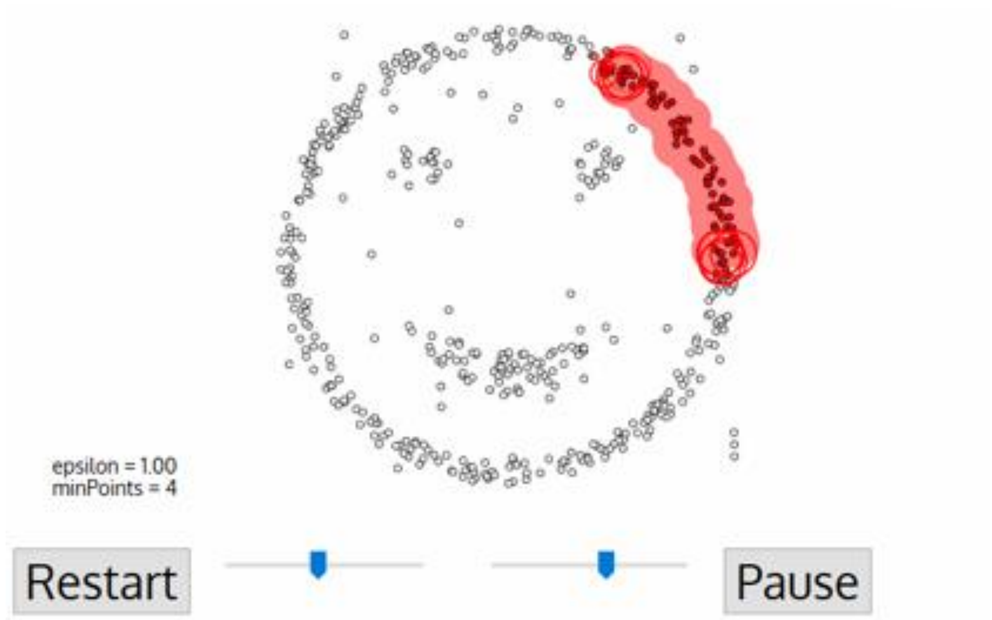
- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.
- A big difference with K-Means is that DBSCAN can define **clusters of arbitrary shape**, not only convex sets.

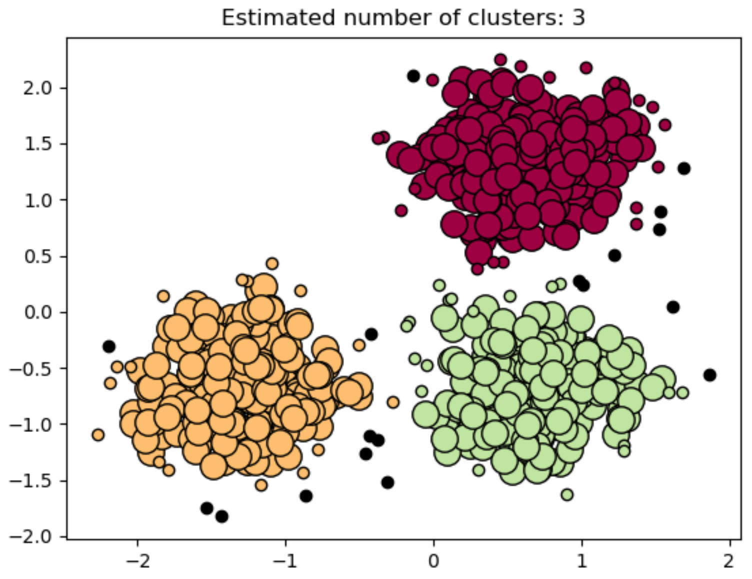




DBSCAN

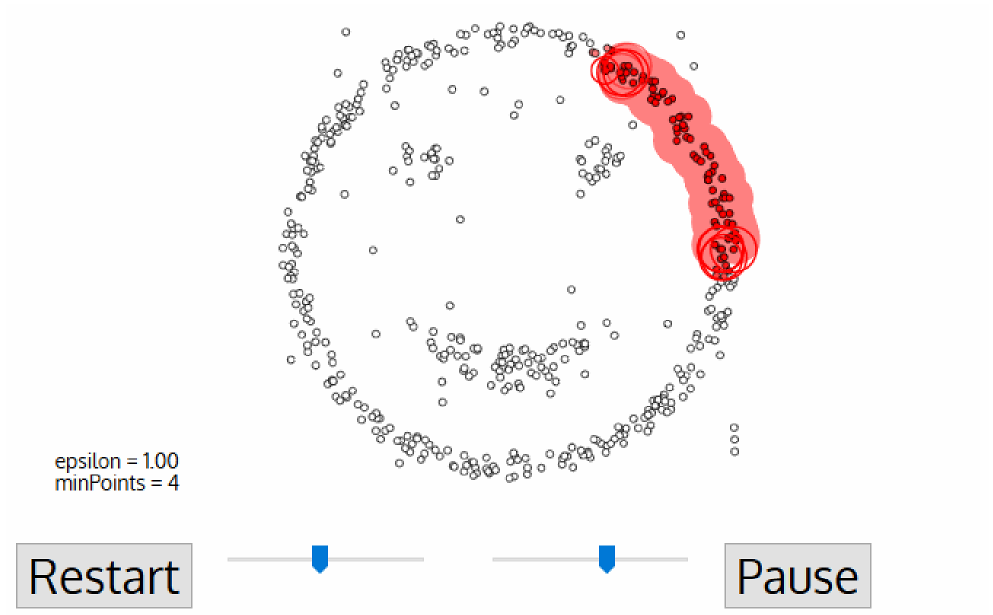
- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.
- A big difference with K-Means is that DBSCAN can define **clusters of arbitrary shape**, not only convex sets.
- The result is **deterministic** (so no randomness involved), but the order of the samples has an impact on the final result. A common strategy for overcoming it consists of performing several random shuffles, repeat the experiments and aggregate the results.

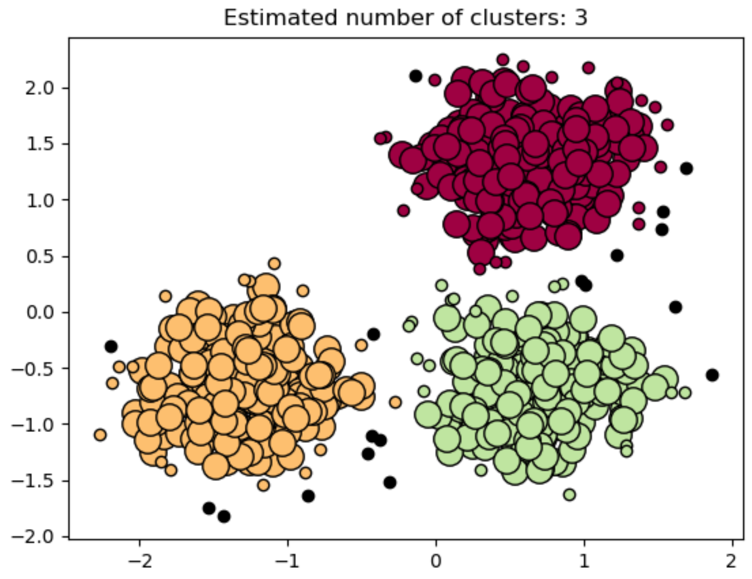




DBSCAN

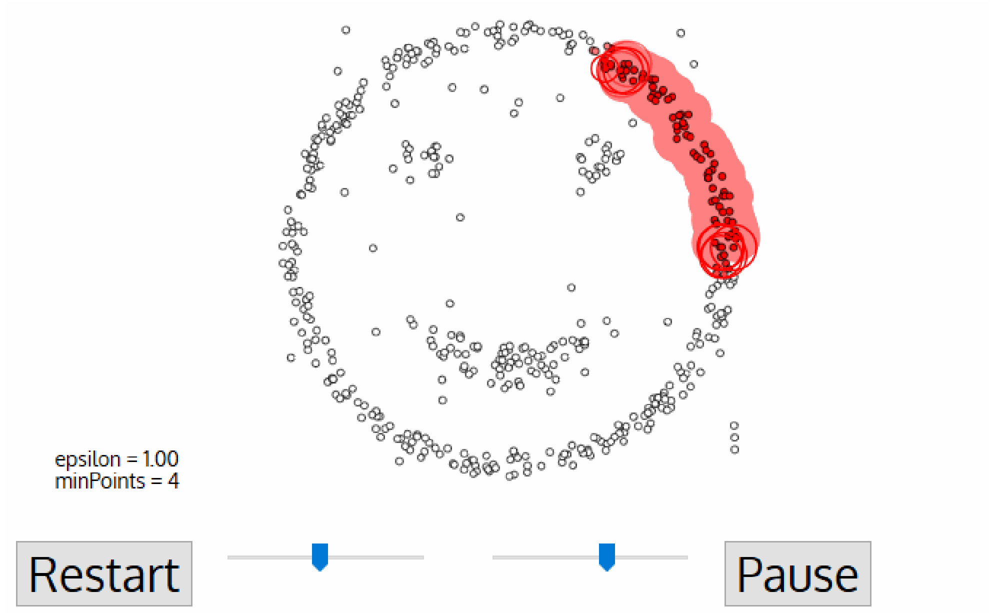
- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.
- A big difference with K-Means is that DBSCAN can define **clusters of arbitrary shape**, not only convex sets.
- The result is **deterministic** (so no randomness involved), but the order of the samples has an impact on the final result. A common strategy for overcoming it consists of performing several random shuffles, repeat the experiments and aggregate the results.

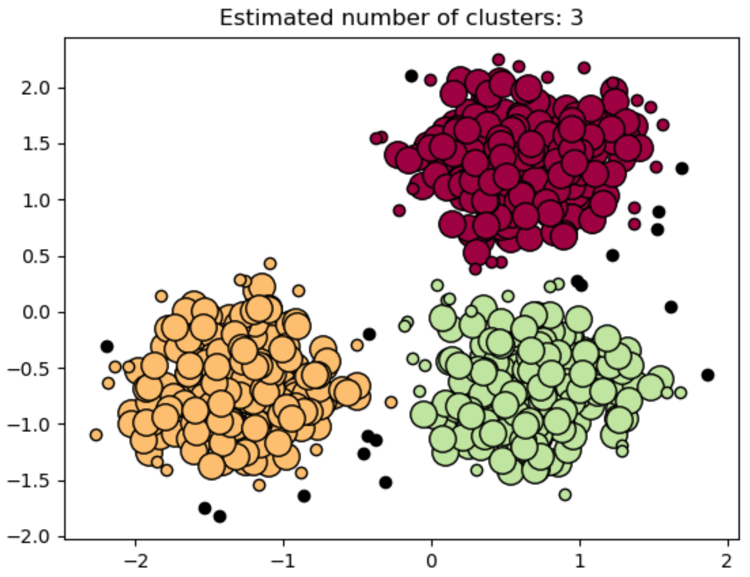




DBSCAN

- No number of clusters is prefixed. Instead, clusters are found understanding them as zones with **high density** separated by zones with **low density**.
- It distinguish between ***core samples*** (in dense zones) y *not-core samples* (at a distance higher than certain threshold from a *core sample)*.
- A big difference with K-Means is that DBSCAN can define **clusters of arbitrary shape**, not only convex sets.
- The result is **deterministic** (so no randomness involved), but the order of the samples has an impact on the final result. A common strategy for overcoming it consists of performing several random shuffles, repeat the experiments and aggregate the results.
- It is **computationally expensive**, but there are variants in the form of **speed-up versions**





Gaussian Mixture Models

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.

Gaussian Mixture Models

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.

Gaussian Mixture Models

Mondragon Unibertsitatea
Escuela Politécnica Superior

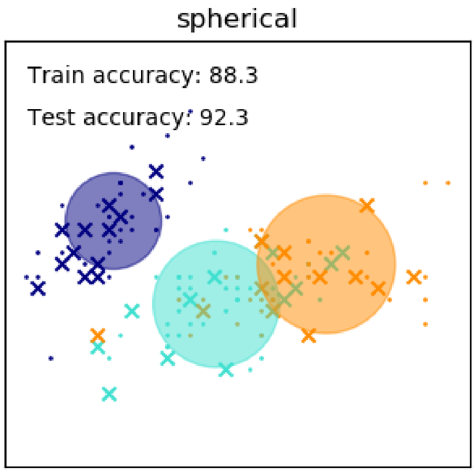
- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.

Gaussian Mixture Models

- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization*** **(EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.
- Depending on the type of covariance we consider, there are 4 variants

Gaussian Mixture Models

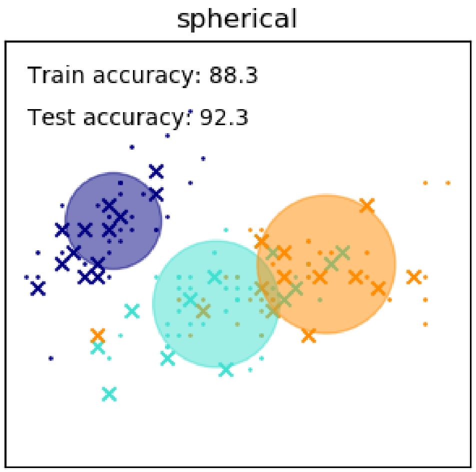
- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.
- Depending on the type of covariance we consider, there are 4 variants:
  - Spherical: Each component has its own individual variance

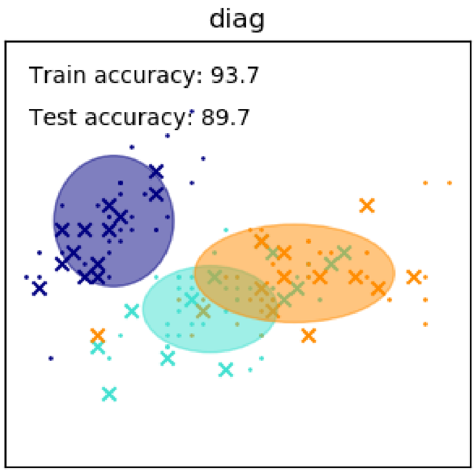


Gaussian Mixture Models

Mondragon Unibertsitatea
Escuela Politécnica Superior

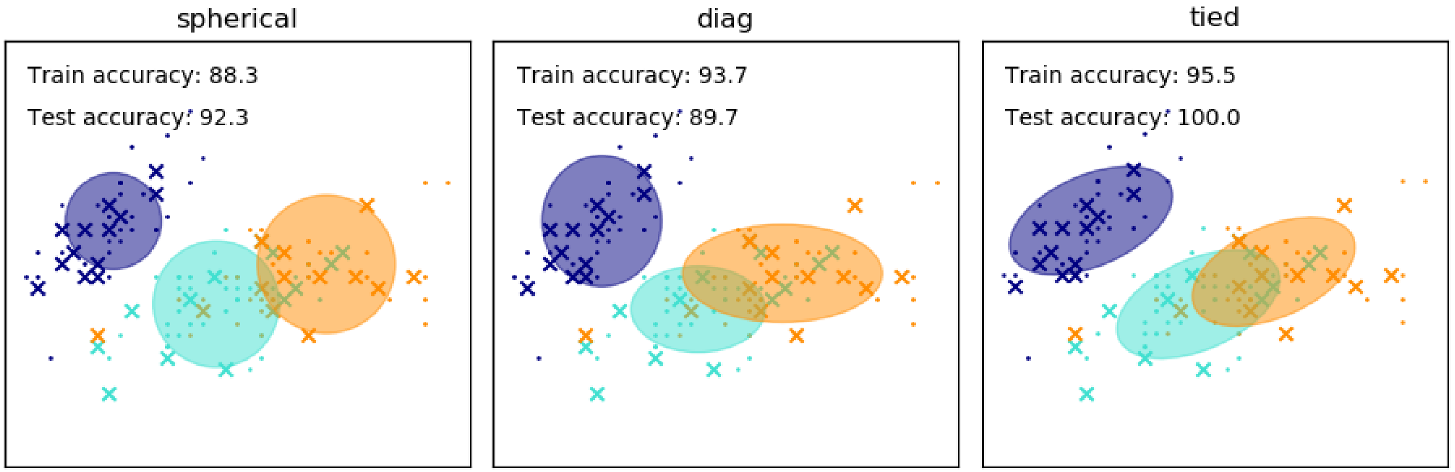
- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.
- Depending on the type of covariance we consider, there are 4 variants:
  - Spherical: Each component has its own individual variance
  - Diagonal: Each component has its own diagonal covariance matrix





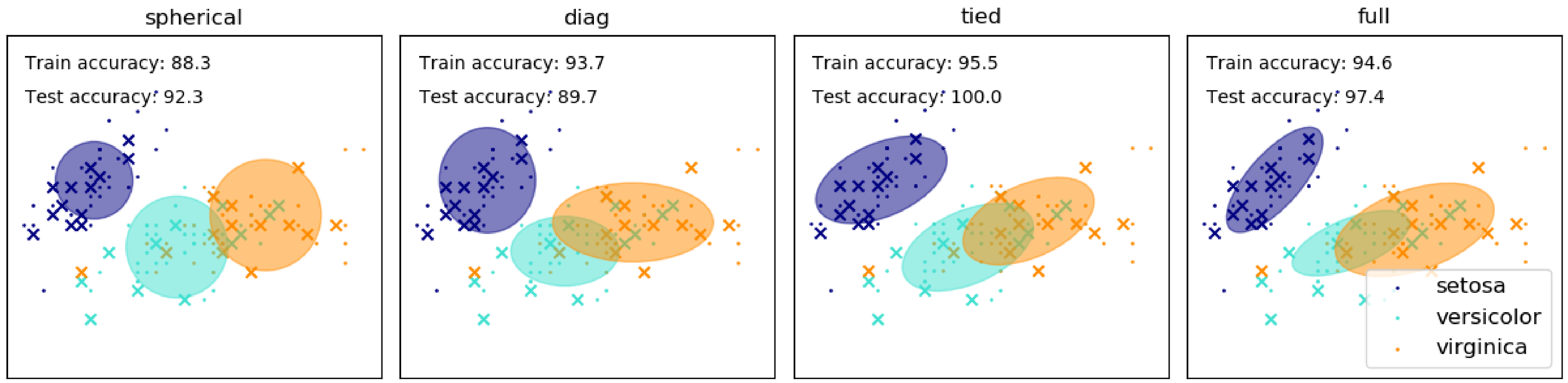
Gaussian Mixture Models

- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.
- Depending on the type of covariance we consider, there are 4 variants:
  - Spherical: Each component has its own individual variance
  - Diagonal: Each component has its own diagonal covariance matrix
  - Tied: All components have the same covariance matrix



Gaussian Mixture Models

- It is a **probabilistic model**. It assumes that all the points have been generated by the **mixture of a finite number of Gaussian distributions** with unknown parameters, which will be approximated by means of the ***expectation maximization* (EM)** algorithm.
- It could be understood as an extension/generalization of K-Means, where the calculations of the distances are conditioned by the data structure.
- Reversely, K-means is equivalent to EM algorithm using a diagonal covariance matrix whose diagonal elements are small and equal.
- Depending on the type of covariance we consider, there are 4 variants:
  - Spherical: Each component has its own individual variance
  - Diagonal: Each component has its own diagonal covariance matrix
  - Tied: All components have the same covariance matrix
  - Full: Each component has its own covariance matrix



Python functions

Mondragon Unibertsitatea
Escuela Politécnica Superior

Algorithm	Function	Use / Parameters of interest

K-Means	KMeans
	MiniBatchKMeans
		sklearn.cluster.KMeans(n_clusters=8, init=‘k-means++’, max_iter=300, tol=0.0001, random_state=None, n_jobs=None)
		sklearn.cluster.MiniBatchKMeans(n_clusters=8, init=‘k-means++’, max_iter=100, batch_size=100, tol=0.0, random_state=None)

Affinity propagation	AffinityPropagation	sklearn.cluster.AffinityPropagation(damping=0.5, max_iter=200, convergence_iter=15, preference=None, affinity=’euclidean’)
Mean shift	MeanShift	sklearn.cluster.MeanShift(bandwidth=None, min_bin_freq=1, cluster_all=True, n_jobs=None)
Hierarchical	AgglomerativeClustering	sklearn.cluster.AgglomerativeClustering(n_clusters=2, affinity=’euclidean’, connectivity=None, linkage=’ward’)
DBSCAN	DBSCAN	sklearn.cluster.DBSCAN(eps=0.5, min_samples=5, metric=’euclidean’, algorithm=’auto’, n_jobs=None)
Gaussian Mixture Models	GaussianMixture	sklearn.mixture.GaussianMixture(n_components=1, covariance_type=’full’, tol=0.001, max_iter=100, init_params=’kmeans’, random_state=None)

3

Validation

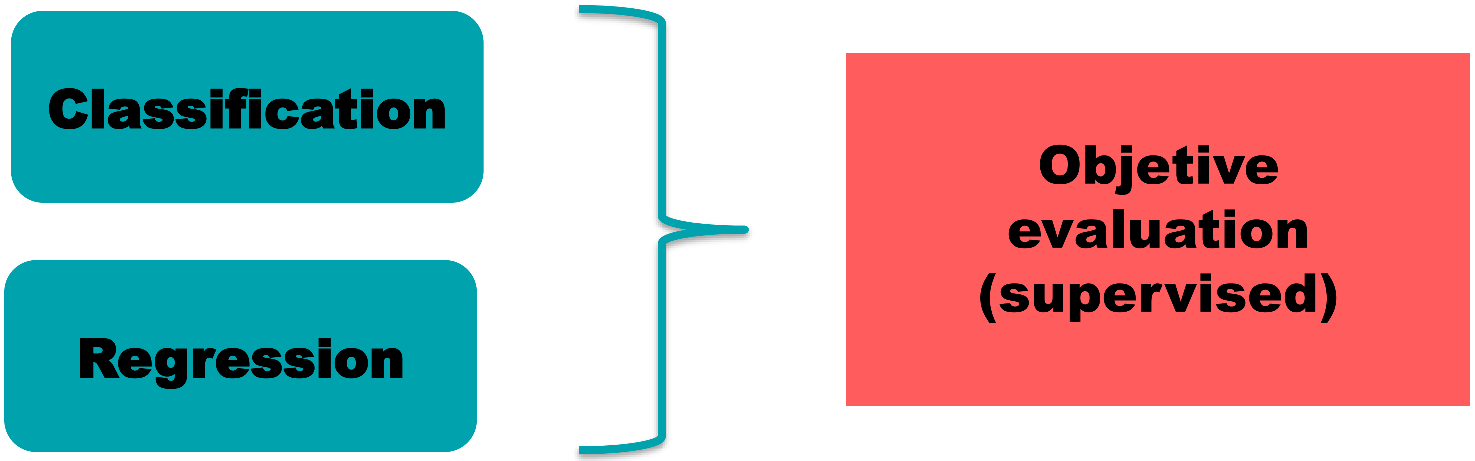
Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

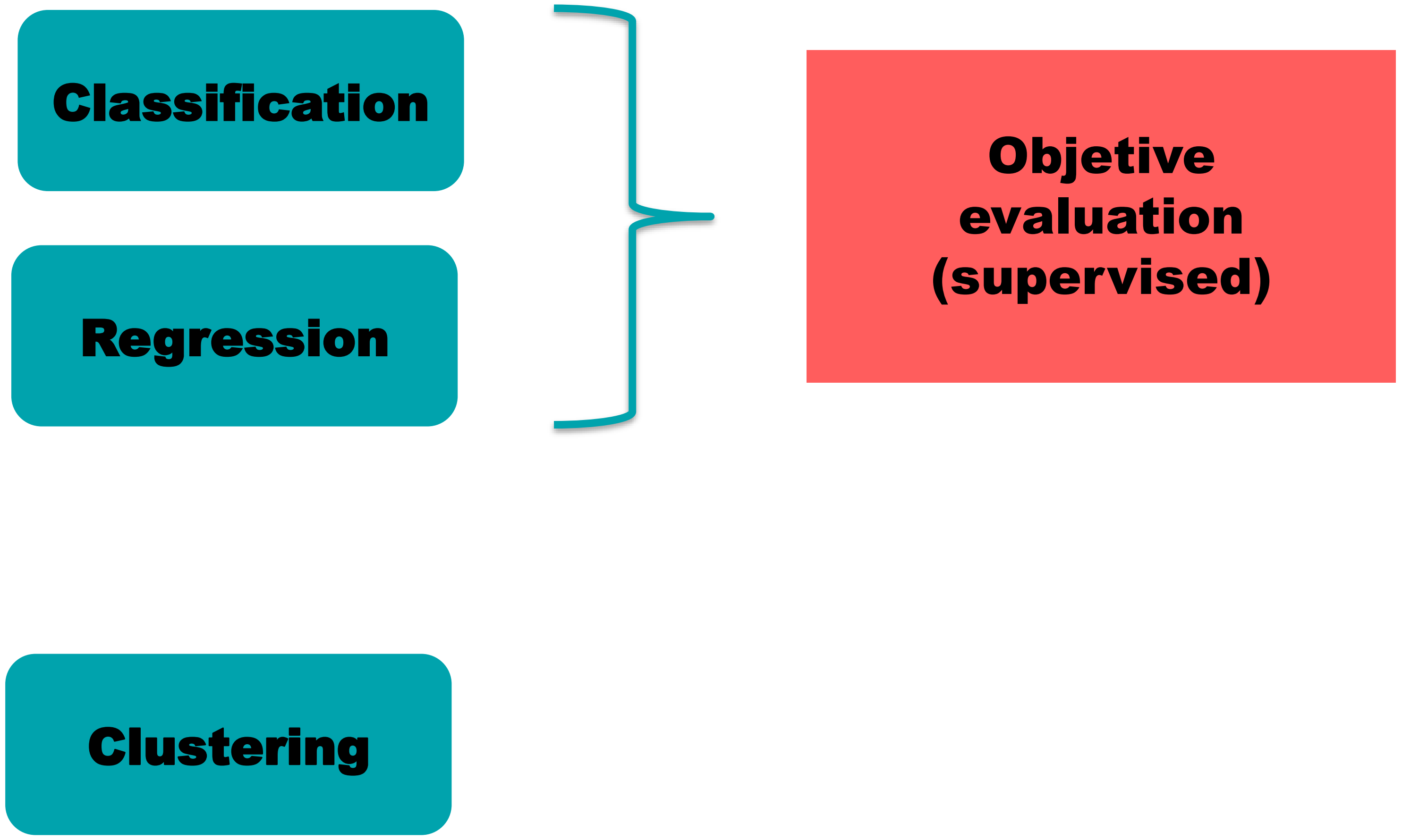
Classification

Regression

Validation


Classification
Regression
Objective evaluation (supervised)

Validation


Classification
Regression
Objetive evaluation (supervised)
Clustering

Validation

Classification

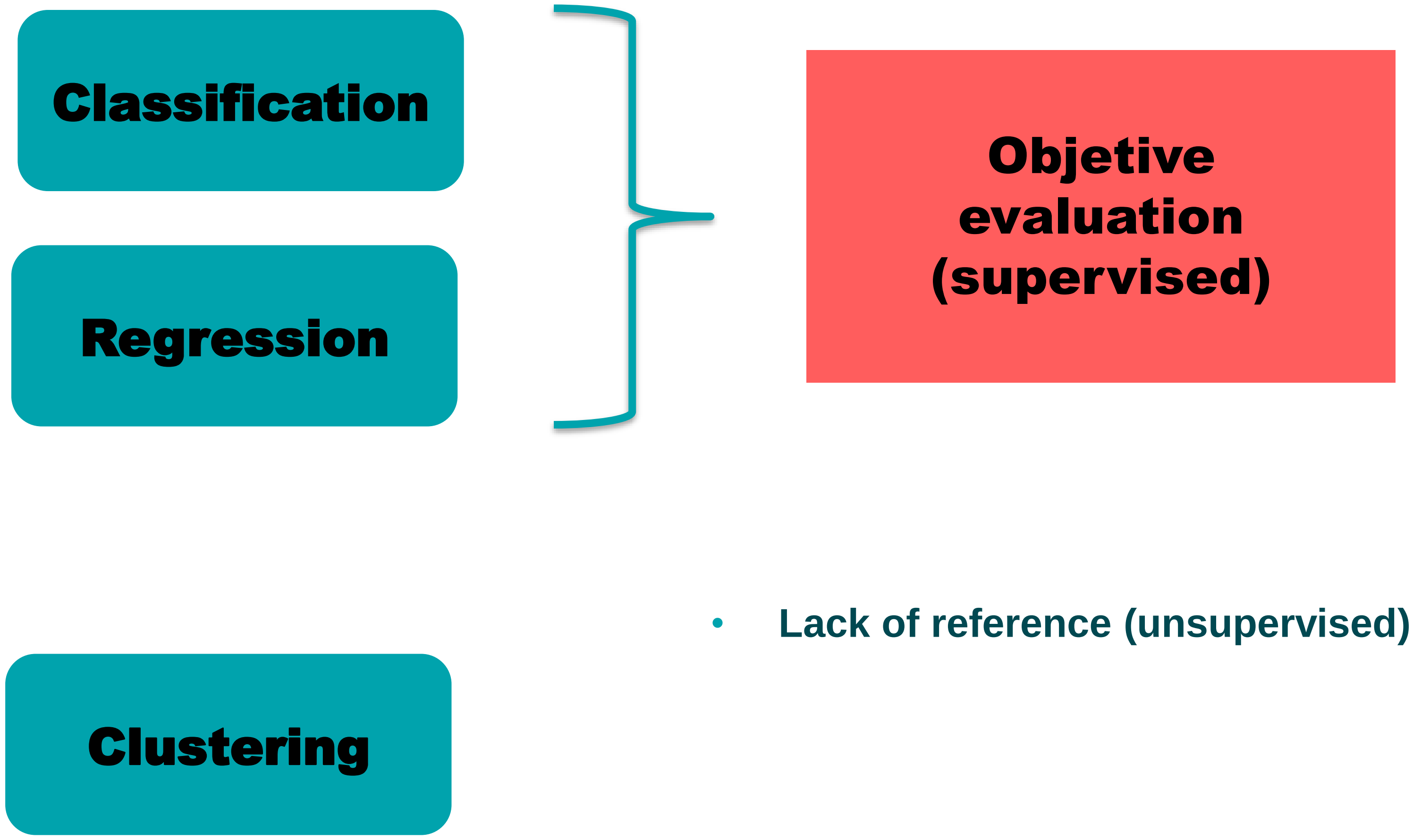
Regression

Objetive evaluation (supervised)

Clustering

?

Validation



- Lack of reference (unsupervised)

Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

Classification

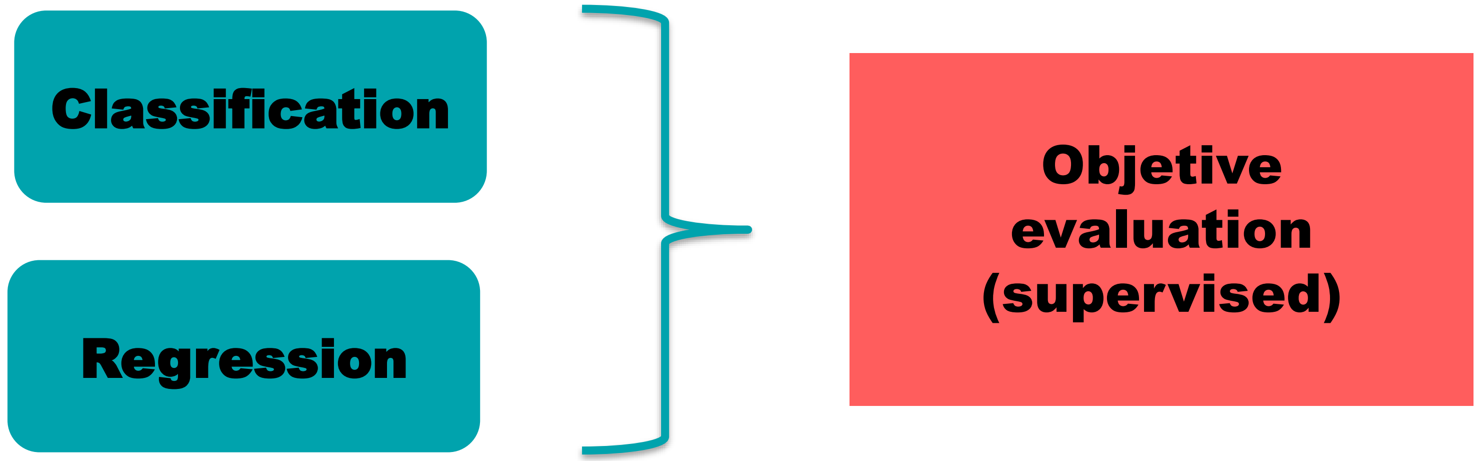
Regression

Objective evaluation (supervised)

Clustering

- Lack of reference (unsupervised)
- Subjetivity of goodness

Validation



Clustering

- Lack of reference (unsupervised)
- Subjetivity of goodness
- Dependency on the similarity measure

Validation

Key questions

Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

Key questions:

- Which is the most adequate **algorithm** for my problem?

Validation

Key questions:

- Which is the most adequate **algorithm** for my problem?
- Which is the right **similarity measure**?

Validation

Key questions:

- Which is the most adequate **algorithm** for my problem?
- Which is the right **similarity measure**?
- If I have to do it by myself, how do I decide the **amount of clusters**?

Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

Key questions:

- Which is the most adequate **algorithm** for my problem?
- Which is the right **similarity measure**?
- If I have to do it by myself, how do I decide the **amount of clusters**?
- How do I evaluate the **quality of a solution** in a clustering problem?

Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

Key questions:

- Which is the most adequate **algorithm** for my problem?
- Which is the right **similarity measure**?
- If I have to do it by myself, how do I decide the **amount of clusters**?
- How do I evaluate the **quality of a solution** in a clustering problem?
- Which **data preprocessing scheme** would contribute to a better result?

Validation

Global answer:

“It depends…”

Validation

Global answer:

“It depends on the chosen way for evaluation.

Usually the right option is searched by a simple trial and error approach”.

Validation

Indices taxonomy

Validation

Indices taxonomy:

- **External indices**: Available external information is used. For instance, if information about classes is known, an (unsupervised) clustering algorithm can be evaluated checking how the underlying class distributions are represented by the clusters.

Validation

Mondragon Unibertsitatea
Escuela Politécnica Superior

Indices taxonomy:

- **External indices**: Available external information is used. For instance, if information about classes is known, an (unsupervised) clustering algorithm can be evaluated checking how the underlying class distributions are represented by the clusters.

- **Internal indices**: Only internal information is used. For instance, checking how homogeneous each of the clusters is (*intra-group homogeneity*), and how heterogeneous different clusters are (*inter-group heterogeneity*).

Validation

Indices taxonomy:

- **External indices**: Available external information is used. For instance, if information about classes is known, an (unsupervised) clustering algorithm can be evaluated checking how the underlying class distributions are represented by the clusters.

- **Internal indices**: Only internal information is used. For instance, checking how homogeneous each of the clusters is (*intra-group homogeneity*), and how heterogeneous different clusters are (*inter-group heterogeneity*).

- **Relative indices**: They adapt the previous ones for comparing pairs of solutions for the same clustering problem. They are out of the scope of this lecture.

Adjusted Rand Index

Mondragon Unibertsitatea
Escuela Politécnica Superior

- **External** index, which needs to know the classes of the samples
- It is assumed that belonging to a cluster is equivalent to being in a class
- All "classifications" are compared
- It is a **symmetrical** score, thus it could be used for measuring **consensus**
- It takes values in [-1, 1]
- A **perfect consensus** gives a score 1, and the expectation in a **random clustering** would be 0. Negative or null values are considered bad scores

- Use in scikit-learn:

```
from sklearn.metrics import adjusted_rand_score
ars = adjusted_rand_score(labels_true, labels_pred)
```

Scores related with mutual information

- Three variants:
  - **Mutual Information Score**: Uses the original definition of mutual information. It suffers with high amounts of clusters.
  - **Normalized Mutual Information Score**: Normalizes using the average of both entropies. This normalization does not alleviate the multi-cluster problem.
  - **Adjusted Mutual Information Score**: Normalizes fixing the aforementioned problem.

- **External** index, which needs to know the classes of the samples
- It is assumed that belonging to a cluster is equivalent to being in a class
- All “classifications” are compared using mutual information, which is a simmetric measure
- It is a **symmetrical** score, thus it could be used for measuring **consensus**
- It takes values in (-∞, 1]
- A **perfect consensus** gives a score 1, and the expectation in a **random clustering** would be 0. Negative or null values are considered bad scores

- Uses in scikit-learn:

```
from sklearn import metrics
mis = metrics.mutual_info_score(labels_true, labels_pred)
nmis = metrics.normalized_mutual_info_score(labels_true, labels_pred)
amis = metrics.adjusted_mutual_info_score(labels_true, labels_pred)
```

Homogeneity, completeness and v-measure

Mondragon Unibertsitatea
Escuela Politécnica Superior

- Two properties, three scores:
  - **Homogeneity**: Each cluster contains members of one single class.
  - **Completeness**: All members of a class are in the same cluster.
  - **V-measure**: Armonic mean of both, which combines both in a balanced way.

- **External** index, which needs to know the classes of the samples
- It is assumed that belonging to a cluster is equivalent to being in a class
- Only v-measure is a **symmetrical** score, thus it is the only one to be used for measuring **consensus**
- Homogeneity and completeness take values in [0, 1], being the higher the better.
- A **perfect consensus** gives a v-score 1. The expectation in a **random clustering** can take different values.
- Obs: For low number of samples or many clusters, *adjusted* type is the right choice.

- Uses in scikit-learn:

```
from sklearn import metrics
hs = metrics.homogeneity_score(labels_true, labels_pred)
cs = metrics.completeness_score(labels_true, labels_pred)
vms = metrics.v_measure_score(labels_true, labels_pred)
hcvm = metrics.homogeneity_completeness_v_measure(labels_true, labels_pred)
```

Fowlkes-Mallows Index

Mondragon Unibertsitatea
Escuela Politécnica Superior

- **External** index, which needs to know the classes of the samples

- They are defined:
  - **True Positive(TP)**: Number of pairs of points belonging to the same clusters, both in the true and the predicted values
  - **False Positive (FP)**: Number of pairs of points belonging to the same clusters in the true values, but not in the predicted ones
  - **False Negative (FN)**: Number of pairs of points belonging to the same clusters in the predicted values, but not in the true ones

- The Fowlkes-Mallows Index is defined as the geometric mean of precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

- It takes values in [0, 1], being the higher the better. In the case of **random clustering** the expectation would be close to 0. An exact 0 value would imply purely independent classifications, and 1 would be an indication of both classifications being the same (except permutations).

- Use in scikit-learn:

```
from sklearn.metrics import fowlkes_mallows_score
fms = fowlkes_mallows_score(labels_true, labels_pred)
```

Contingency matrix

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [‘a’, ‘a’, ‘a’, ‘b’, ‘b’, ‘b’]$$
$$label\_pred = [’D’, ’D’, ’E’, ’E’, ’F’, ’F’]$$
$$cm = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$$

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\boxed{‘a’, ‘a’, ‘a’}, ‘b’, ‘b’, ‘b’]$$
$$label\_pred = [′D′, ′D′, ′E′, ′E′, ′F′, ′F′]$$
$$cm = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$$

Of the 3 true samples in ‘a’

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\boxed{'a', 'a', 'a',} 'b', 'b', 'b']$$
$$label\_pred = ['D', 'D', 'E', 'E', 'F', 'F']$$
$$cm = \begin{pmatrix} \boxed{2} & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$$

Of the 3 true samples in 'a',
2 have been predicted in 'D'

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\boxed{'a', 'a', 'a',} 'b', 'b', 'b']$$
$$label\_pred = ['D', 'D', 'E', 'E', 'F', 'F']$$
$$cm = \begin{pmatrix} \boxed{2} & \boxed{1} & 0 \\ 0 & 1 & 2 \end{pmatrix}$$

Of the 3 true samples in ‘a’,
2 have been predicted in ‘D’,
1 in ‘E’

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\boxed{‘a’, ‘a’, ‘a’,} ‘b’, ‘b’, ‘b’]$$
$$label\_pred = [’D’, ’D’, ’E’, ’E’, ’F’, ’F’]$$
$$cm = \begin{pmatrix} \boxed{2} & \boxed{1} & \boxed{0} \\ 0 & 1 & 2 \end{pmatrix}$$

Of the 3 true samples in ‘a’,
2 have been predicted in ‘D’,
1 in ‘E’ and 0 in ‘F’

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\text{‘a’}, \text{‘a’}, \text{‘a’}, \boxed{\text{‘b’}, \text{‘b’}, \text{‘b’}}]$$
$$label\_pred = [\text{’D’}, \text{’D’}, \text{’E’}, \text{’E’}, \text{’F’}, \text{’F’}]$$
$$cm = \begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 1 & 2 \end{pmatrix}$$

Of the 3 true samples in ‘b’,
0 have been predicted in ‘D’

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [\text{'a'}, \text{'a'}, \text{'a'}, \boxed{\text{'b'}, \text{'b'}, \text{'b'}}]$$
$$label\_pred = [\text{'D'}, \text{'D'}, \text{'E'}, \text{'E'}, \text{'F'}, \text{'F'}]$$
$$cm = \begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & \boxed{1} & 2 \end{pmatrix}$$

Of the 3 true samples in 'b',
0 have been predicted in 'D',
1 in 'E'

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Contingency matrix

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is an **external measure** (not an index) that precises to know the classes of the samples, which are considered as clusters by themselves
- It includes, for the pairs of clusters (predicted and true), an intersection level
- Example:

$$label\_true = [‘a’, ‘a’, ‘a’, \boxed{‘b’, ‘b’, ‘b’}]$$
$$label\_pred = [’D’, ’D’, ’E’, ’E’, ’F’, ’F’]$$
$$cm = \begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & \boxed{1} & \boxed{2} \end{pmatrix}$$

Of the 3 true samples in ‘b’,
0 have been predicted in ‘D’,
1 in ‘E’ and 2 in ‘F’

- Use in scikit-learn:

```
from sklearn.metrics.cluster import contigency_matrix
cm = contigency_matrix(labels_true, labels_pred)
```

Silhouette Coefficient

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is an **internal index** that only needs the model itself
- They are defined:
  - **a**: Average distance between a sample and the rest of points of the same cluster
  - **b**: Average distance between a sample and the points of the closest cluster
- The Silhouette Coefficient (SC) is defined, for a single sample, as:

$$s = \frac{b - a}{\max(a, b)}$$

- For the whole set, the average SC for all samples is calculated.
- It takes values in [-1, 1], being the higher the better (i.e. clusters being more dense and better separated). Scores close to 0 indicate clusters overlapping
- Obs: If we use SC for comparing different clustering solutions provided by different clustering methods, we should take into account that SC is generally higher for convex clusters. Be careful when comparing clustering solutions coming from density-based methods with others, thus the latter tend to generate non-convex clusters.
- Use in scikit-learn:

```
from sklearn.metrics import silhouette_score
ss = silhouette_score(X, labels, metric='euclidean')      [X contains data (by default)]
ss = silhouette_score(X, labels, metric='precomputed')   [X contains pairwise distances]
```

Calinski-Harabaz Index

- It is an **internal index** that only needs the model itself
- It is also known as **Variance Ratio Criterion**, because it is defined as the ratio between the average variance between clusters and the internal dispersion inside them
- The higher it is, the more dense and disperse clusters are, thus better the clusters collection
- It is fast to calculate
- Obs: If we use it for comparing different clustering solutions provided by different clustering methods, we should take into account that it is generally higher for convex clusters. Be careful when comparing clustering solutions coming from density-based methods with others, thus the latter tend to generate non-convex clusters. Besides, the use of centroids forces us to use Euclidean distancies.
- Use in scikit-learn:

```
from sklearn.metrics import calinski_harabaz_score
dbs = calinski_harabaz_score(X, labels)
```

Davies-Bouldin Index

Mondragon Unibertsitatea
Escuela Politécnica Superior

- It is an **internal index** that only needs the model itself

- It is defined a similarity measure between clusters C_i and C_j, denoted by R_{ij}, which takes into account:
  - s_i: average distance between each point in cluster C_i and its centroid (diameter)
  - d_{ij}: the distance between the centroids of both clusters

- In the original paper, the similarity measure was defined as

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

- Then, the Davies-Bouldin Indexis defined as

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

- The lower the index, the better the clusters collection. The lower possible value is 0.

- It is faster to be calculated than Silhouette Coefficient

- Obs: If we use it for comparing different clustering solutions provided by different clustering methods, we should take into account that it is generally higher for convex clusters. Be careful when comparing clustering solutions coming from density-based methods with others, thus the latter tend to generate non-convex clusters. Besides, the use of centroids forces us to use Euclidean distancies.

- Use in scikit-learn:

```
from sklearn.metrics import davies_bouldin_score
dbs = davies_bouldin_score(X, labels)
```

Mondragon Unibertsitatea

Escuela Politécnica Superior

Eskerrik asko

Muchas gracias

Thank you

Carlos Cernuda
ccernuda@mondragon.edu

MGEP
Goiru, 2
20500 Arrasate – Mondragon
Tlf. 662420414