FOUNDATIONS OF STATISTICAL ANALYSIS & MACHINE LEARNING

Amric Trudel
amric.trudel@epita.fr

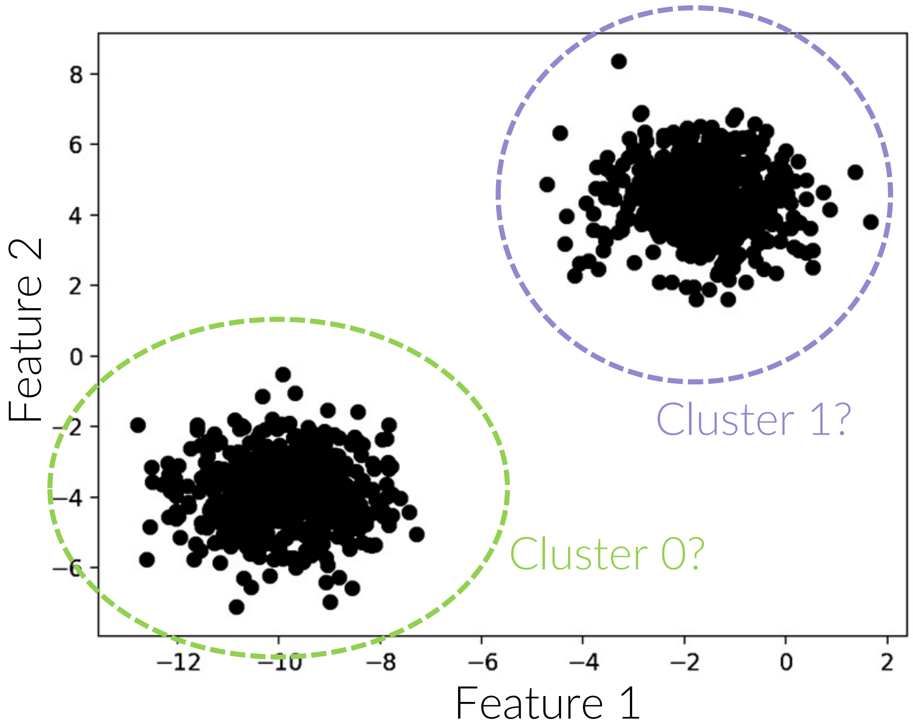
COURSE PROGRAM

Structure

PREPARATION	Data exploration
	Data preprocessing
REGRESSION	Linear regression with one variable
	Multiple and polynomial regression
CLASSIFICATION	Logistic regression
	Classification model assessment
	k-NN, Decision Tree, SVM
CLUSTERING	**k-means, hierarchical clustering**
DIMENSIONALITY REDUCTION	Principal Components Analysis
ALL NOTIONS	Final assignment

CLUSTERING

Problem statement



Unsupervised

- No associated responses to check
- Unknown number of clusters

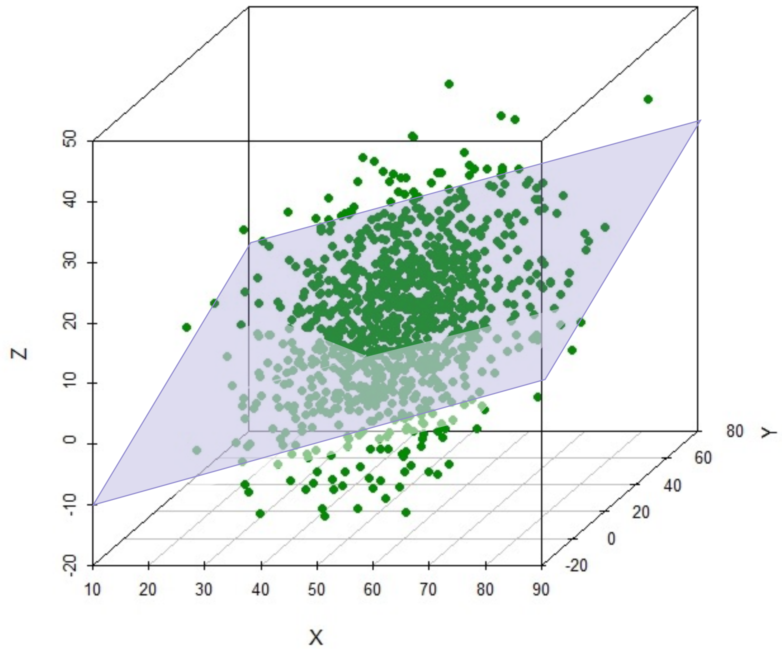
CLUSTERING

General approach

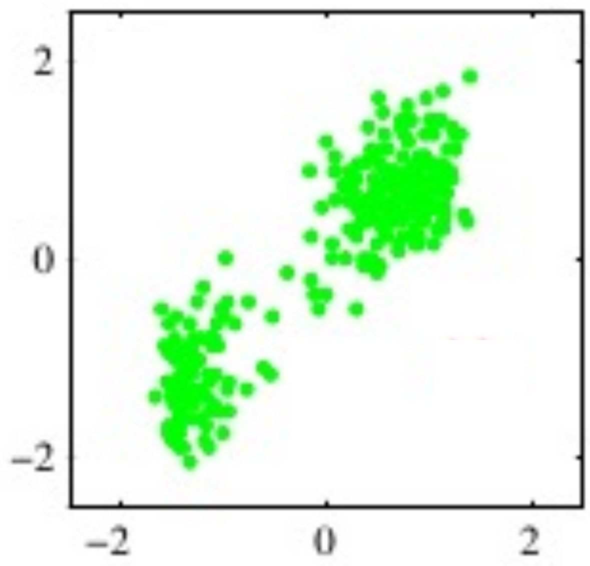
1. Get some intuition from **data inspection** (dimension reduction, visualization, etc.)
2. **Choose a model**
3. **Fine-tune** the model based on a cost function

CLUSTERING

Data inspection



Dimensionality reduction



Visualization

CLUSTERING

Model choice

- k-Means
- Hierarchical Clustering
- Gaussian Mixtures
- Density-based Clustering
- …

CLUSTERING

Model fine-tuning

- Iterative process (on the data set, on the number of clusters, etc.)
- Cost functions:
  - Intra-cluster proximity to center
  - Inter-cluster distance
  - Likelihood
  - Intra-cluster density

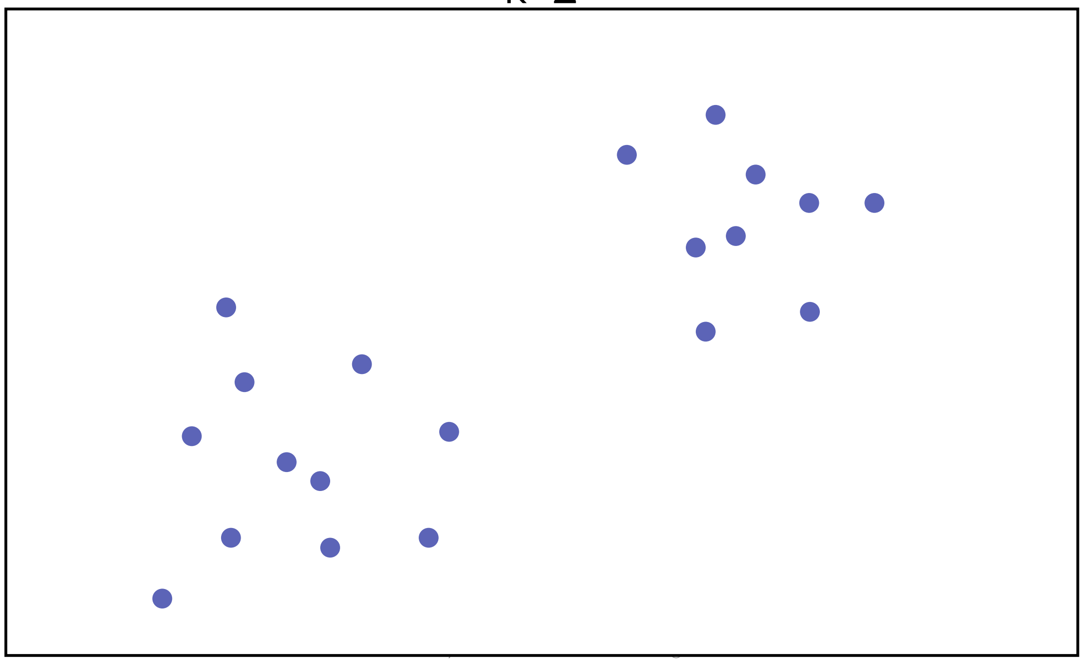
- The cost functions can be used for model comparison

K-MEANS

Principle

- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).

k=2

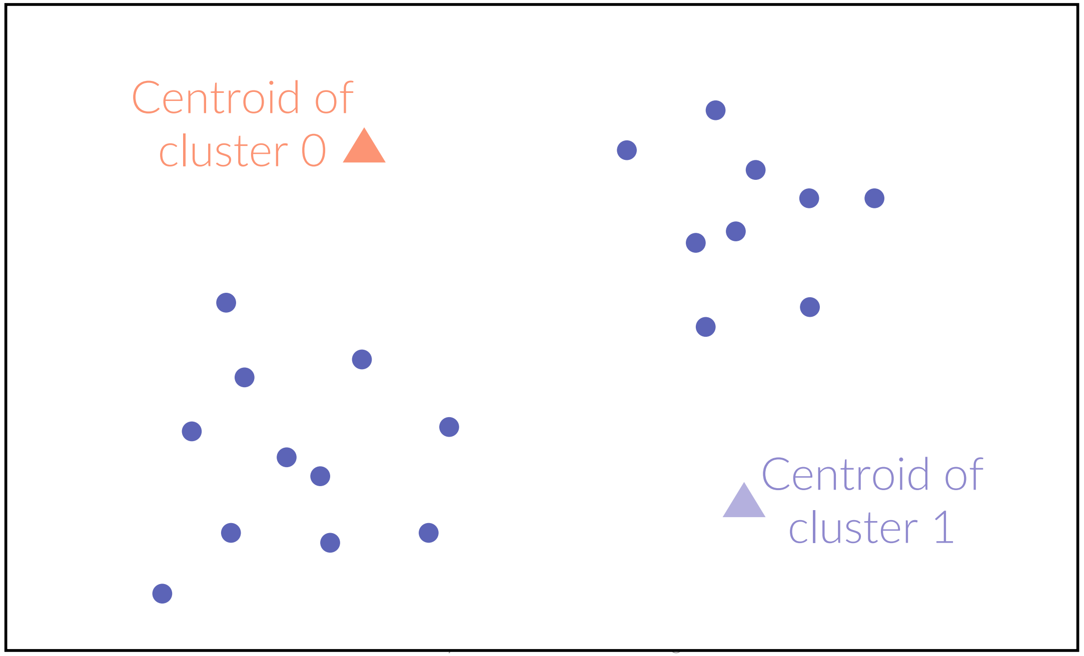


K-MEANS

Principle

- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).

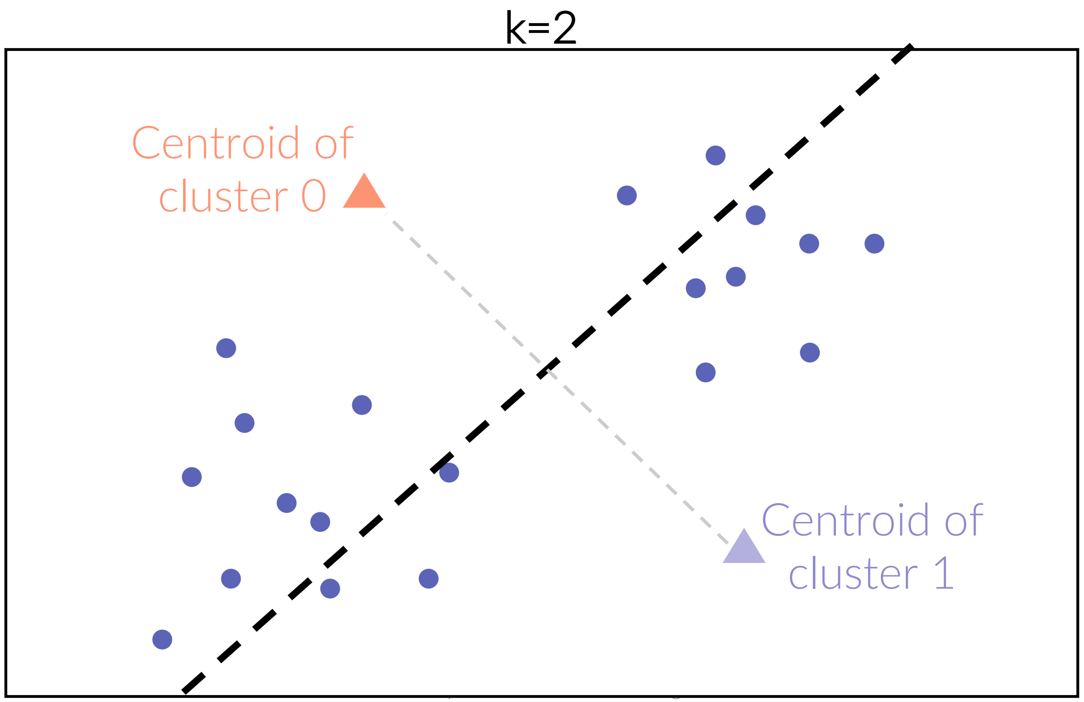
k=2



K-MEANS

Principle

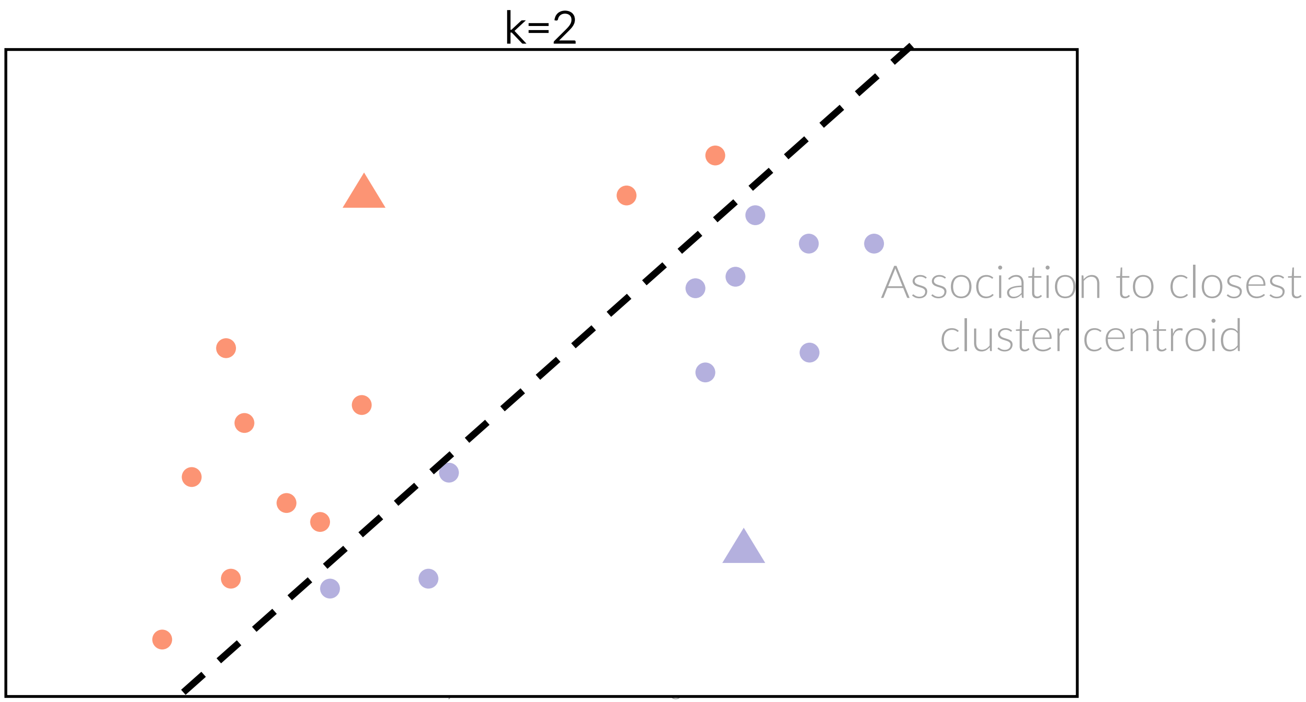
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

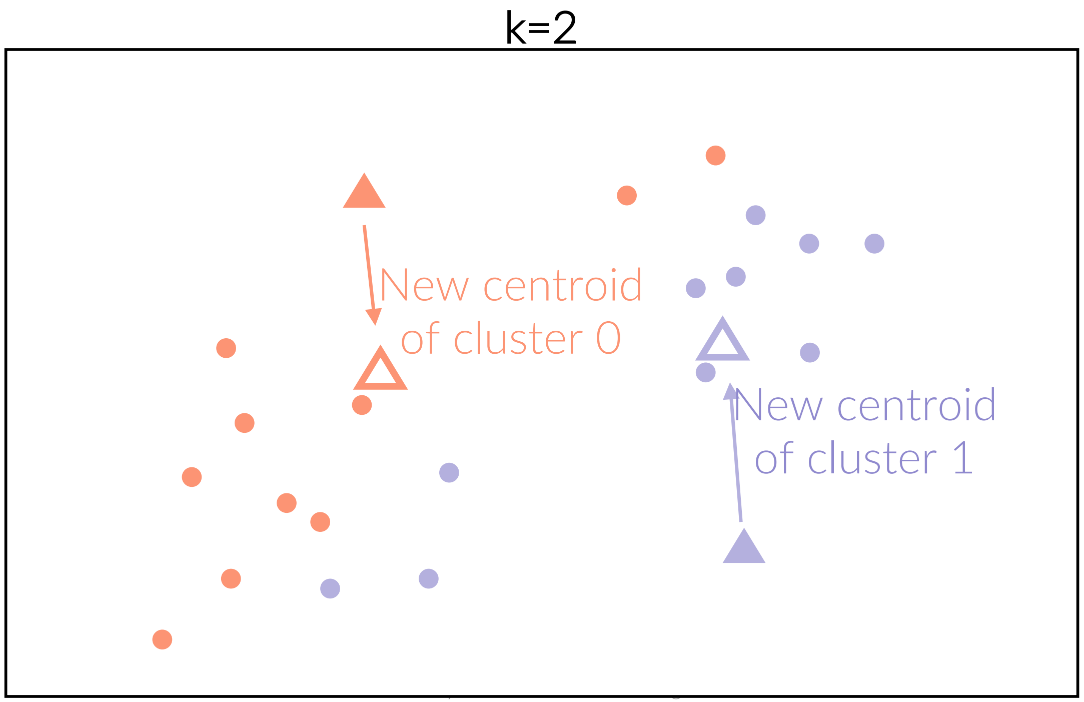
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

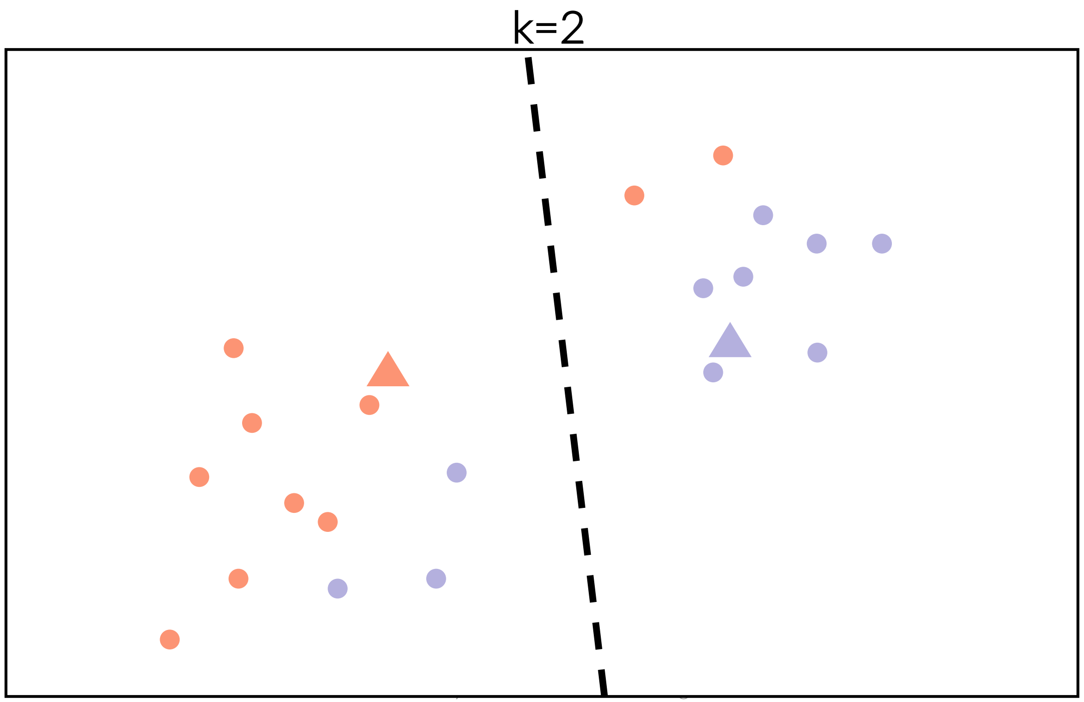
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

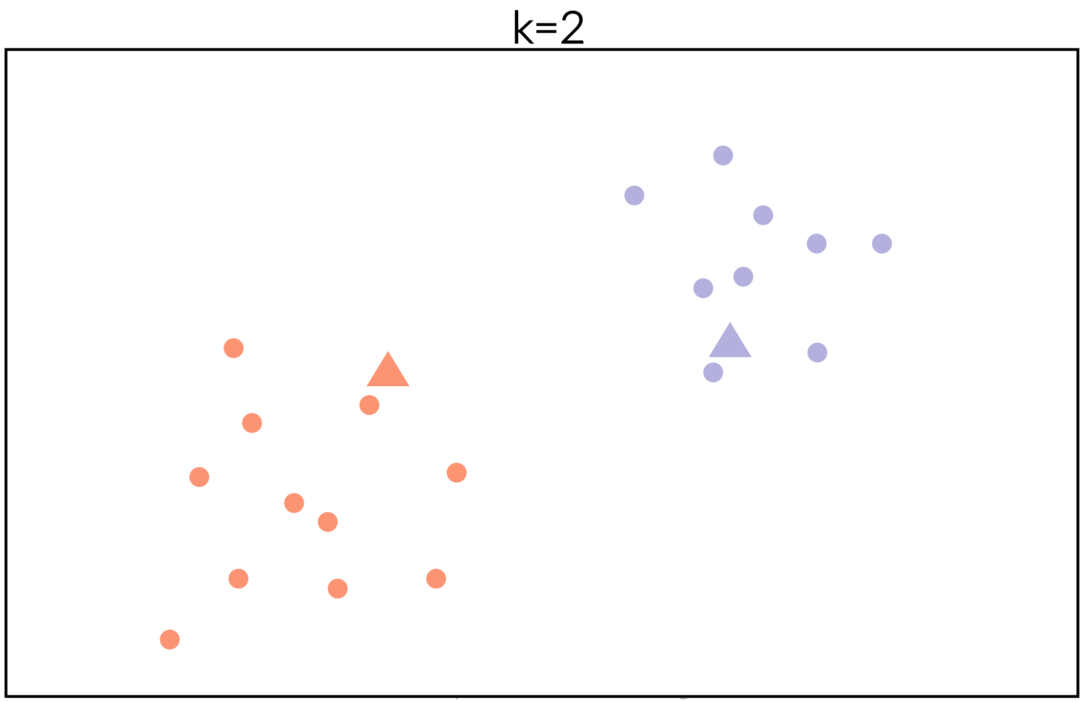
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

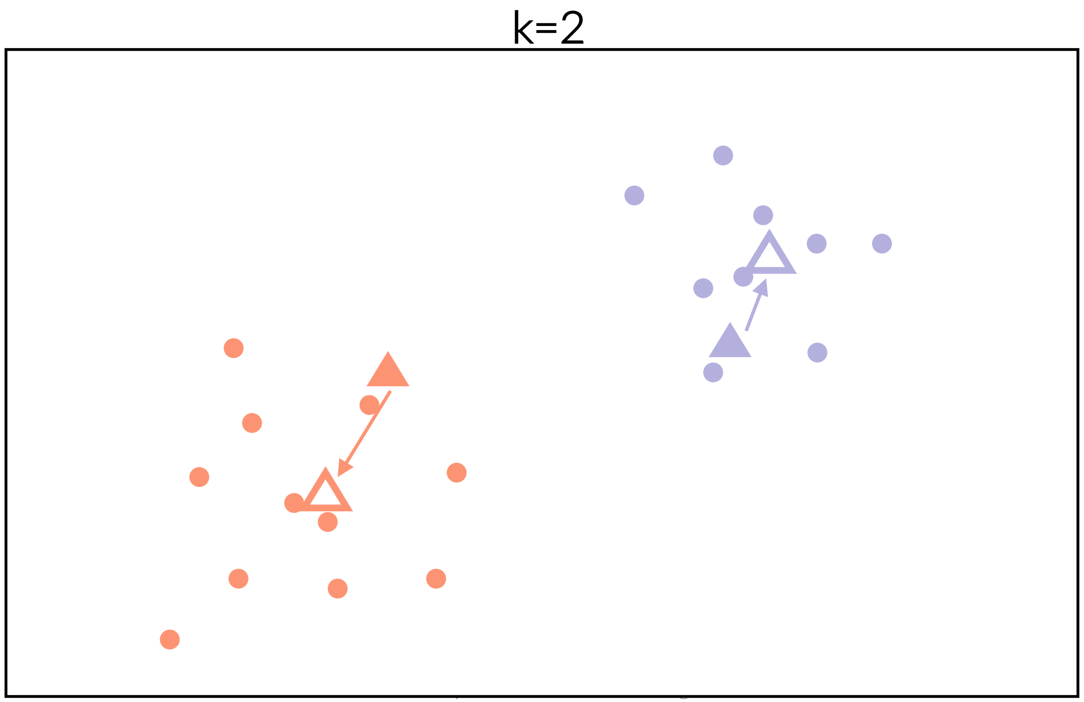
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

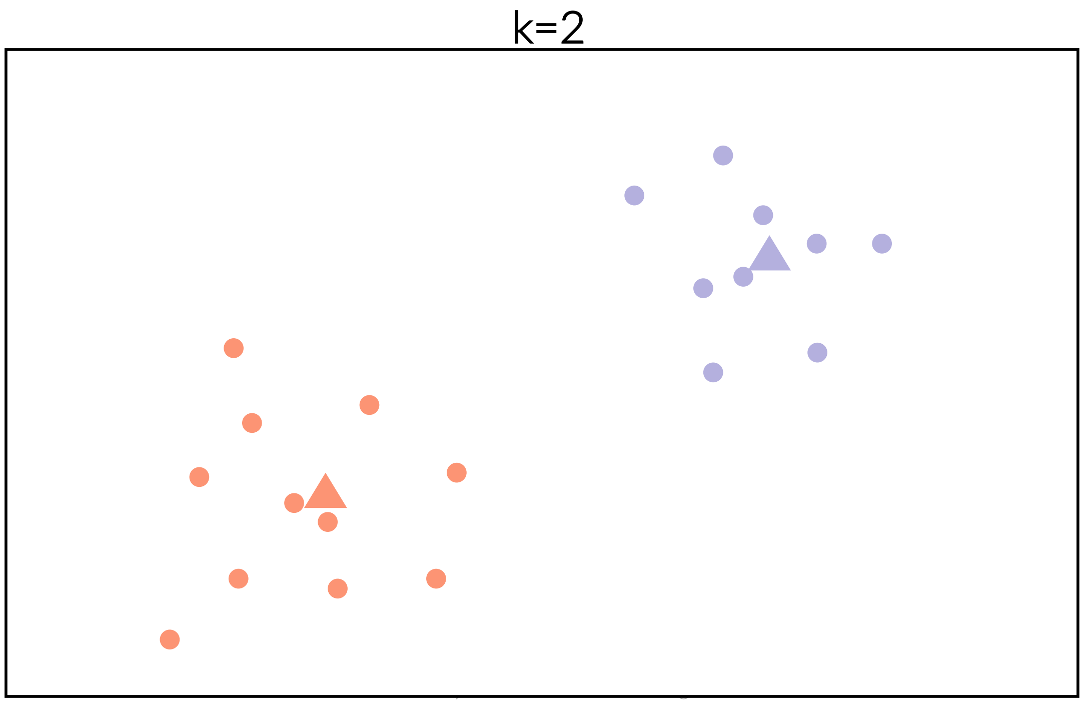
- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

Principle

- Objective: find k subgroups in the dataset that minimize the intra-cluster distances (homogeneity) and maximize the inter-cluster distances (separation).



K-MEANS

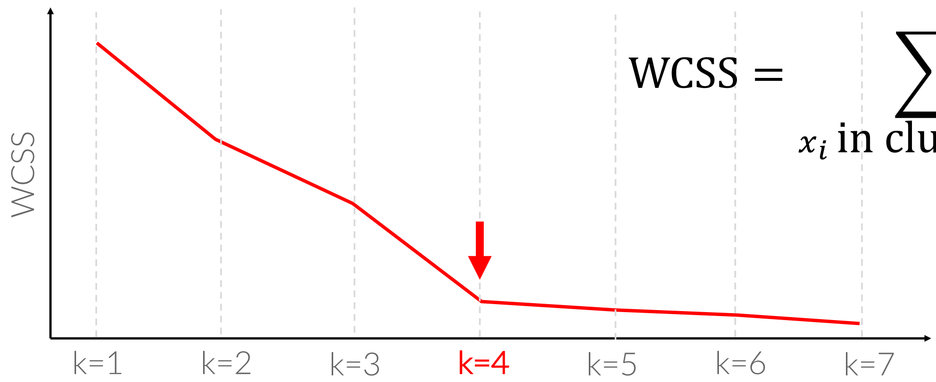
Process

- Choose the number k of clusters
- Attribute random positions to the k centroids
- Assign each data point to the closest centroid
- Recalculate the position of each centroid
- Repeat steps 3 and 4 until the centroids do not change position

K-MEANS

Process

- Choice of distance: Euclidian, etc.
- Choice of k: e.g. through the Elbow Method:
  - Compute the final **WCSS (within-cluster sums of squares) - a.k.a inertia** distances between each point and its centroid for increasing values of k
  - Stop increasing k when it stops providing significative WCSS reduction



$$\text{WCSS} = \sum_{x_i \text{ in cluster } 1} \text{dist}(x_i, c_1)^2 + \sum_{x_i \text{ in cluster } 2} \text{dist}(x_i, c_2)^2 + \cdots$$

K-MEANS

Python implementation

- Training a k-Means model for clustering:

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 5)
kmeans.fit(X)
```

- Predicting cluster attribution:

```
y_pred = kmeans.predict(X)
```

- Getting the coordinates of the cluster centers:

```
kmeans.cluster_centers_
```

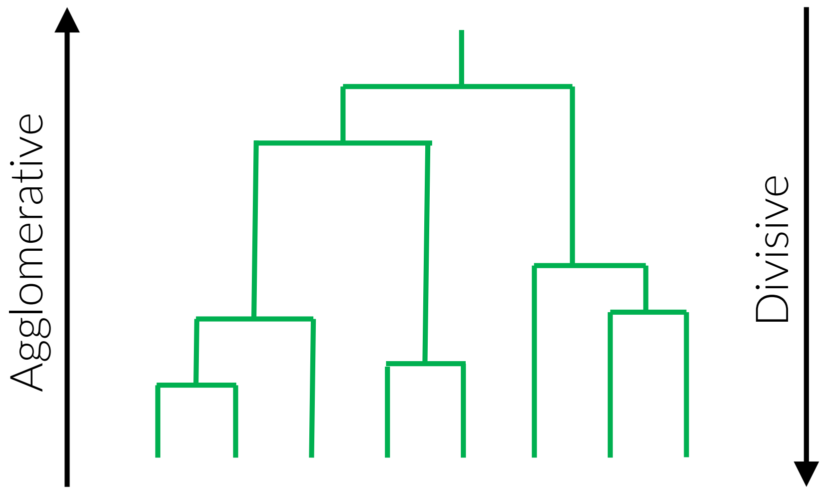
- Getting the WCSS of that model:

```
kmeans.inertia_
```

HIERARCHICAL CLUSTERING

Principle

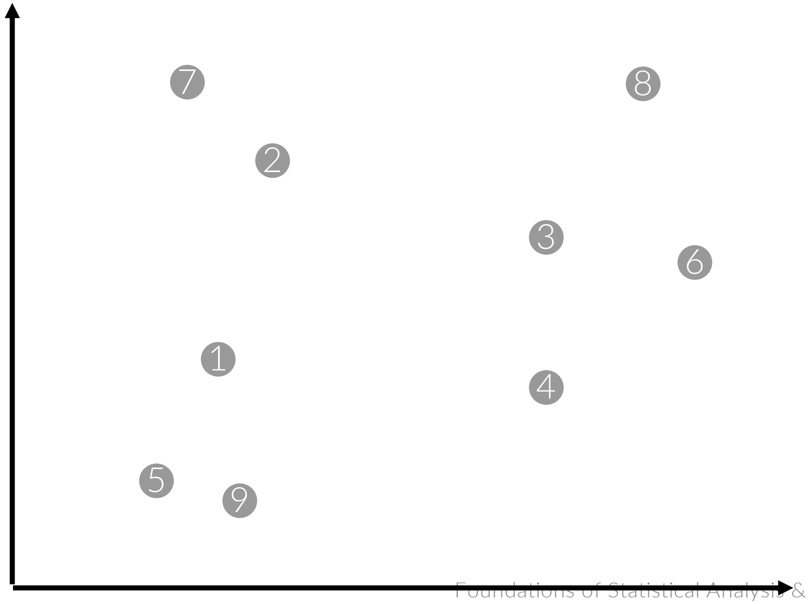
- Construction of hierarchical clusters



HIERARCHICAL CLUSTERING

Principle

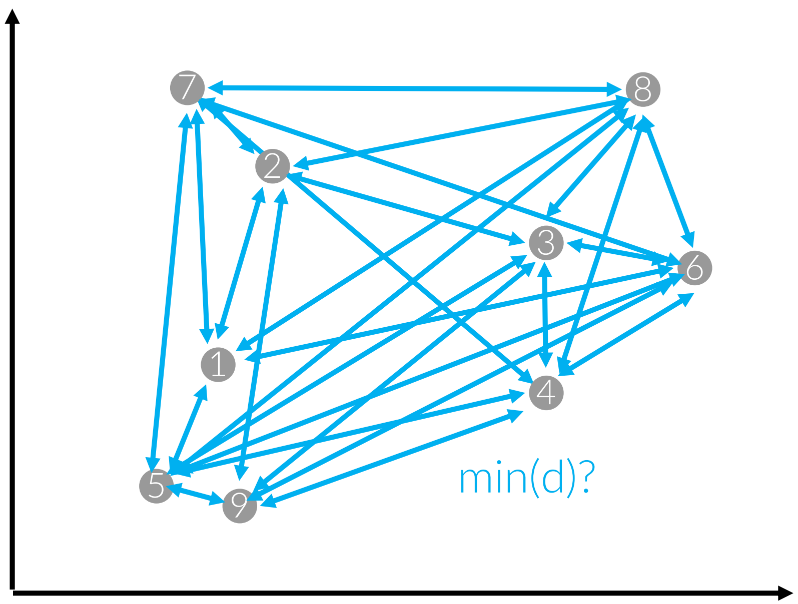
- Dendrogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

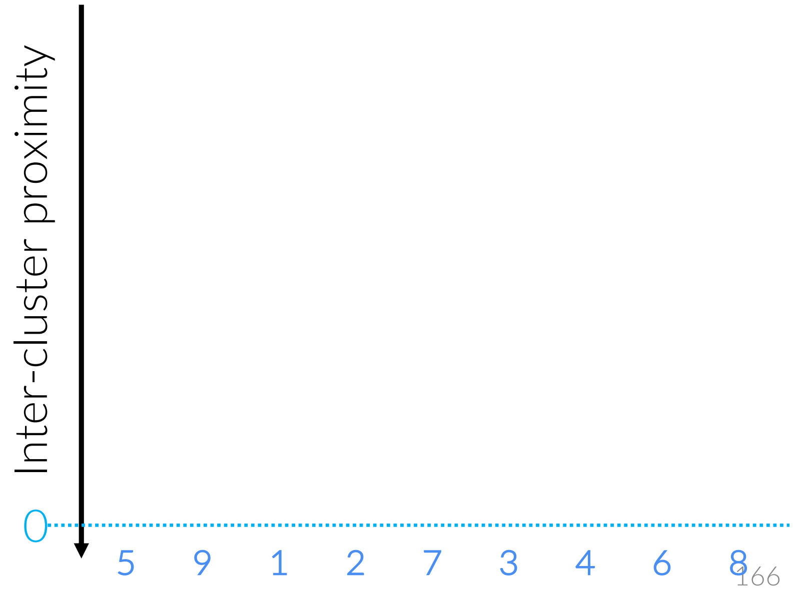


HIERARCHICAL CLUSTERING

Principle

- Dendrogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

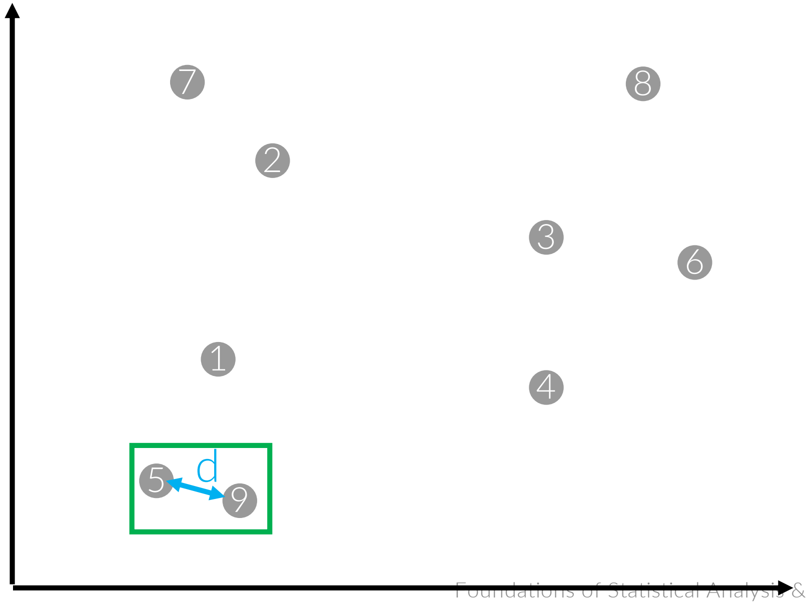


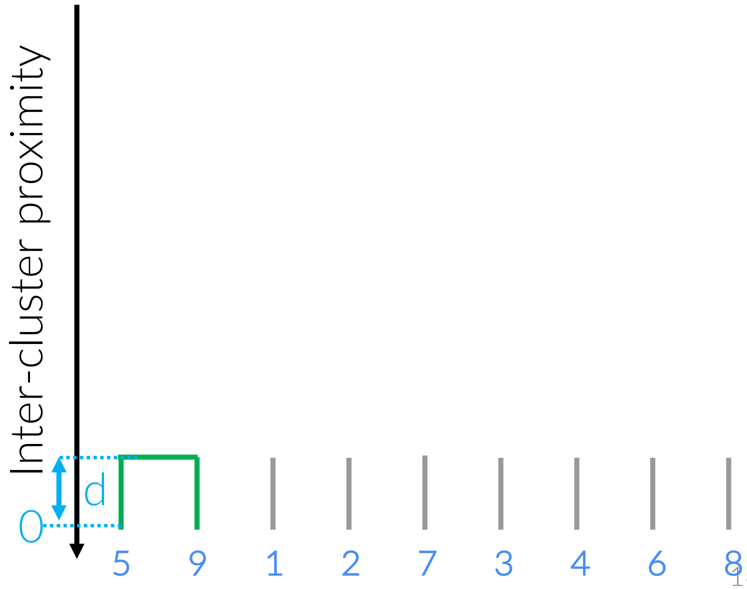


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

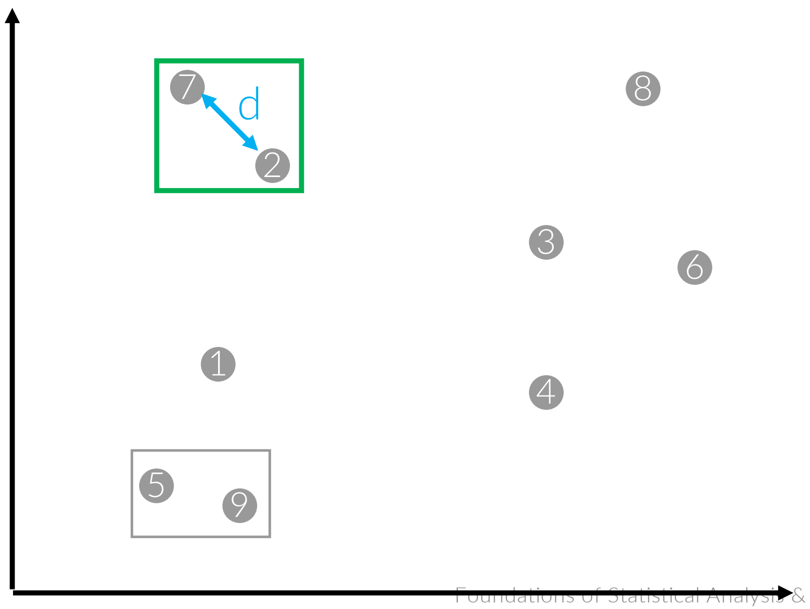


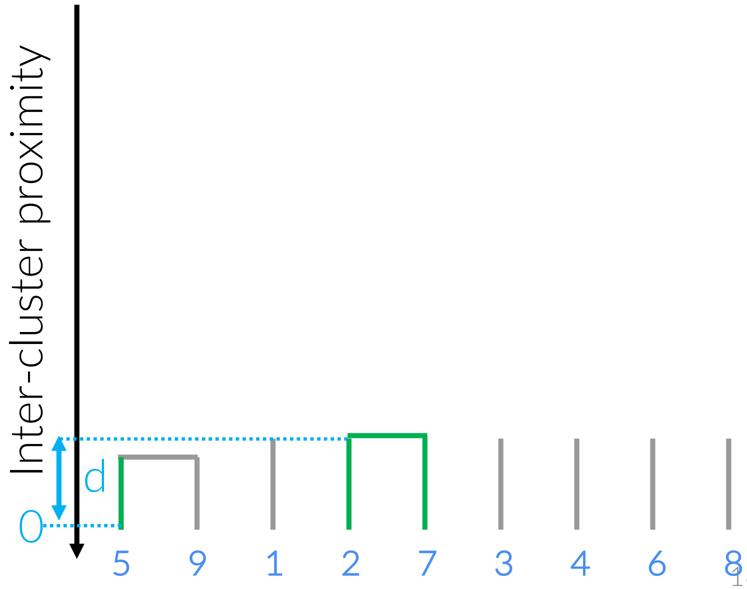


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

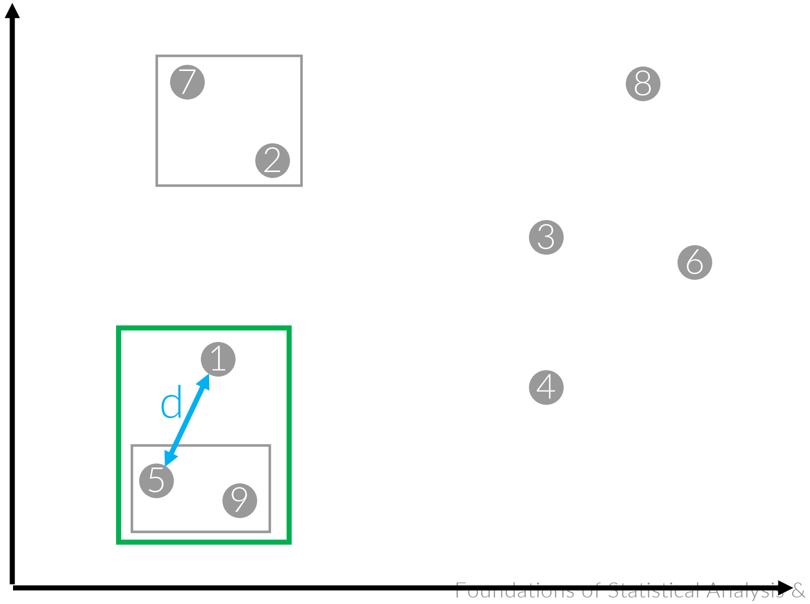


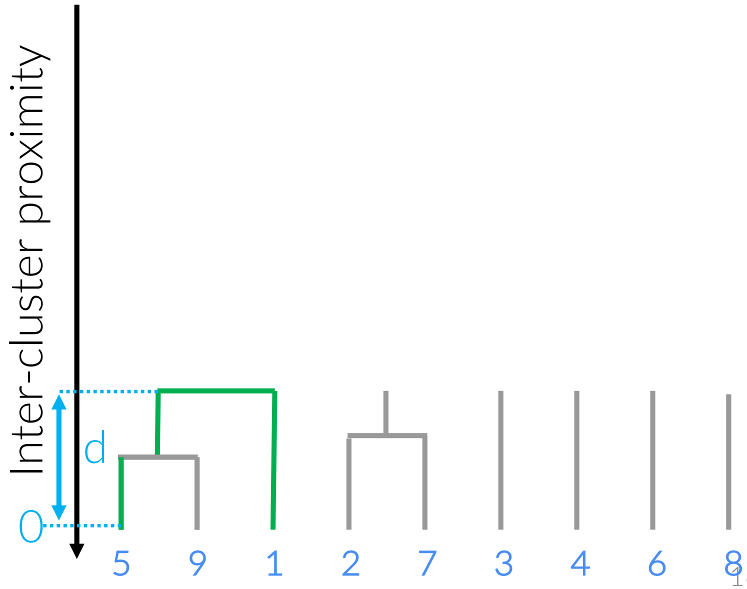


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

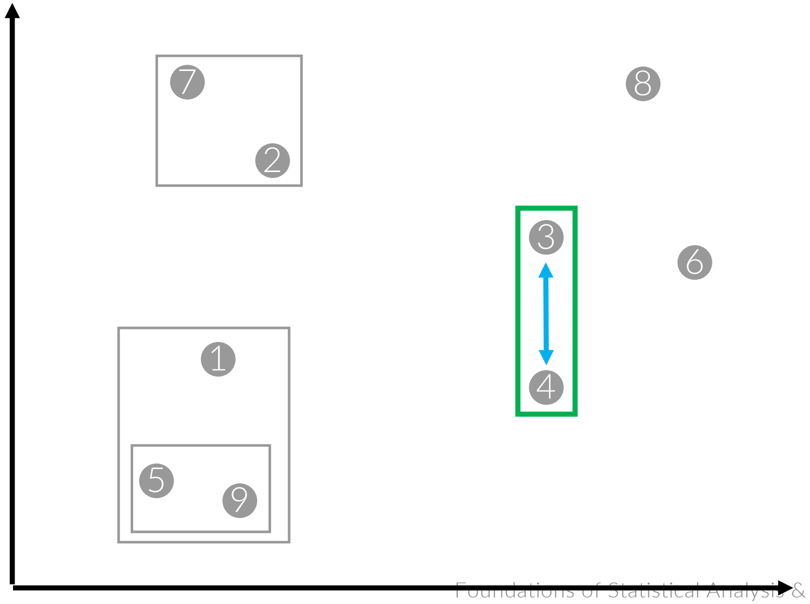


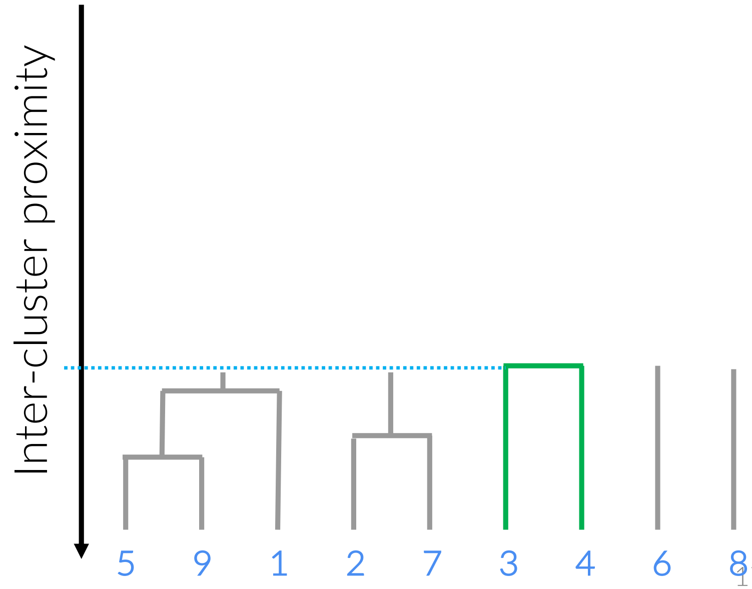


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

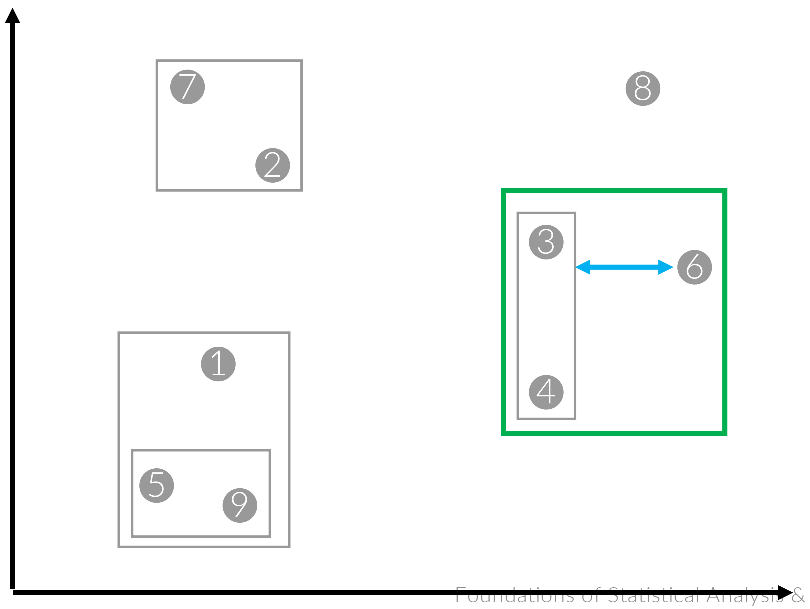


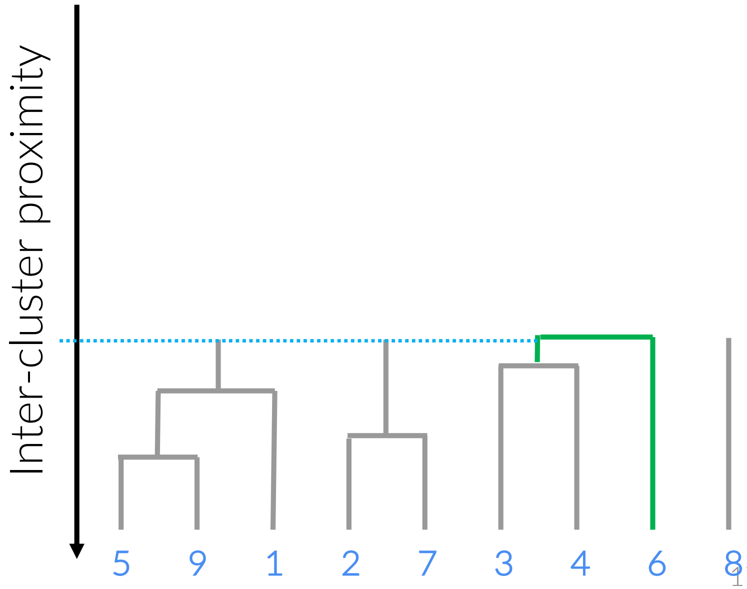


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

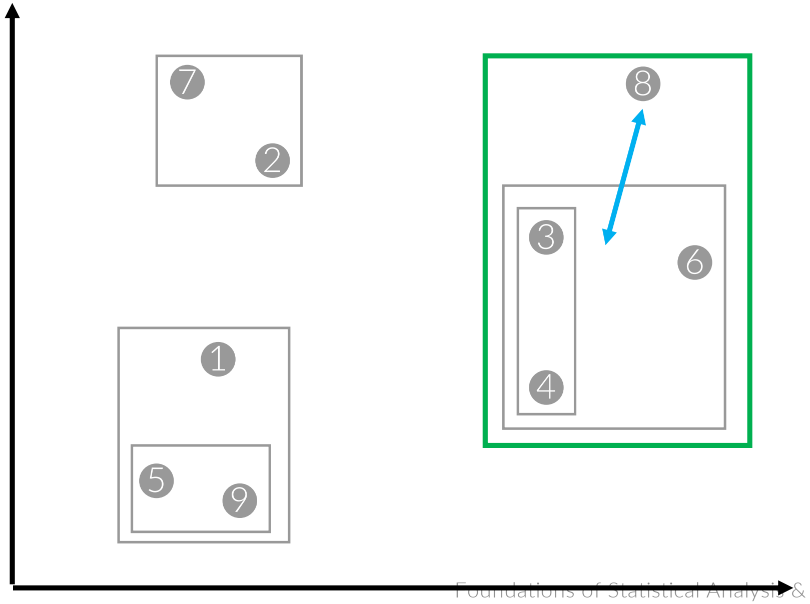


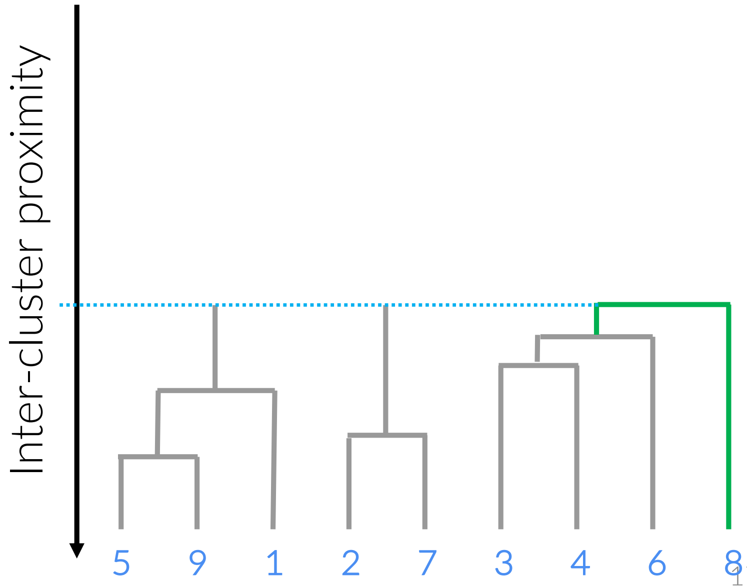


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

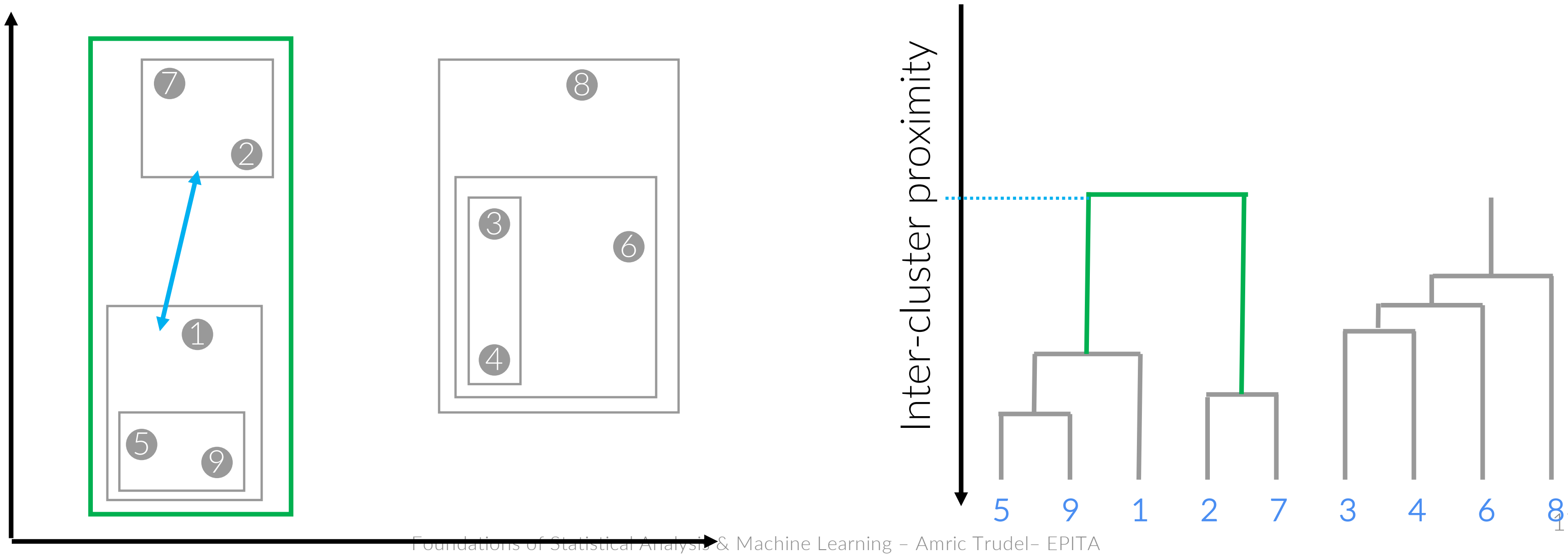




HIERARCHICAL CLUSTERING

Principle

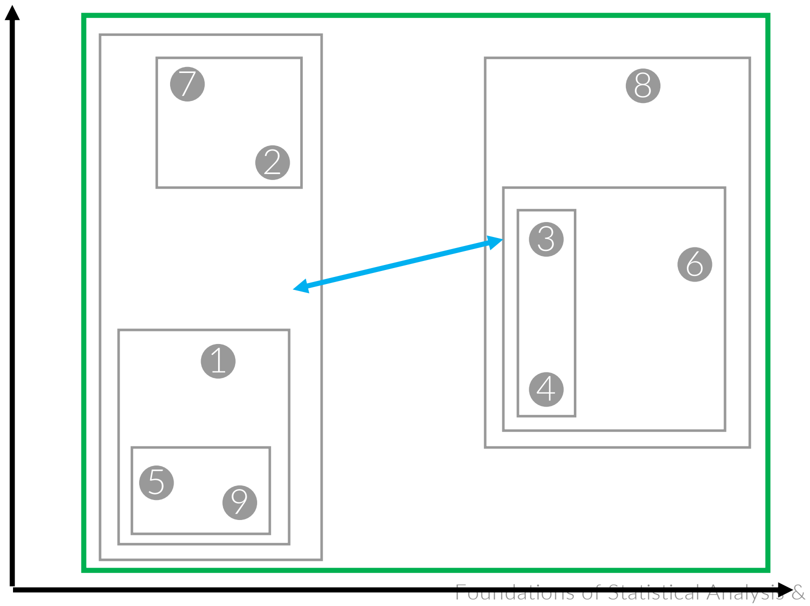
- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

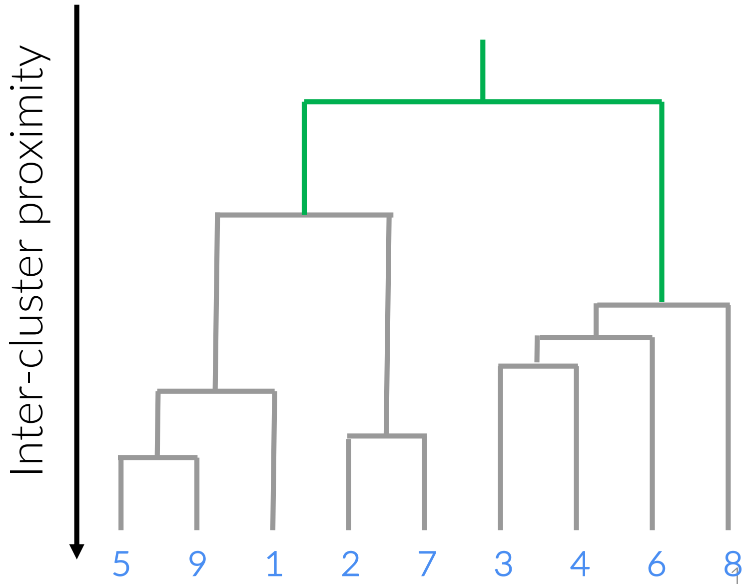


HIERARCHICAL CLUSTERING

Principle

- Dendogram (agglomerative): starting from the leaves (the data points) and combining clusters up to the trunk
- Criteria of cluster similarity/proximity

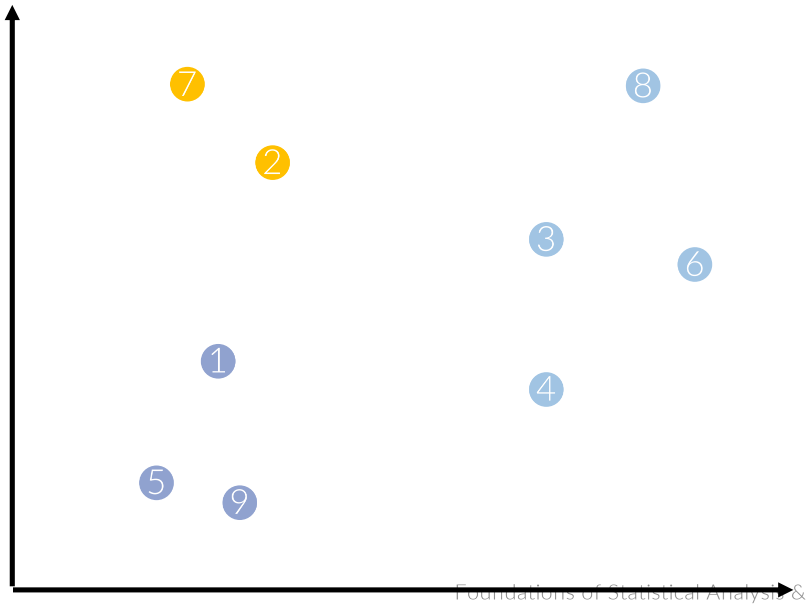


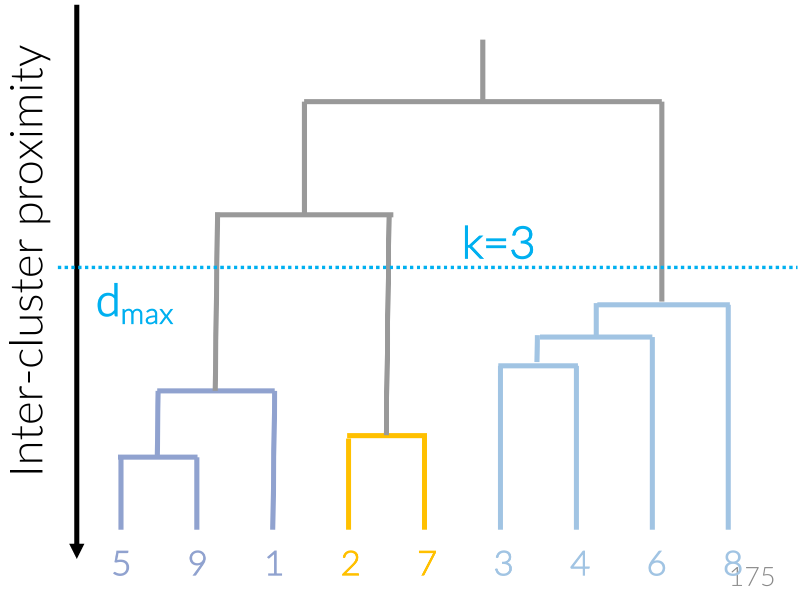


HIERARCHICAL CLUSTERING

Principle

- Horizontal cut to define k hierarchical clusters





HIERARCHICAL CLUSTERING

Process

- Build the dendrogram:
  - Create one cluster for each data point
  - Compute the proximity matrix of the distances between each pair of clusters
  - Merge the two closest clusters
  - Update the proximity matrix
  - Repeat the two previous steps until only a single cluster remains
- Make a horizontal cut across the dendrogram (max value of inter-cluster distance). The distinct sets of observation beneath the cut can be interpreted as clusters.

HIERARCHICAL CLUSTERING

Process

- Distance between clusters:
  - Single linkage: minimum of the distances between all observations of the two sets
  - Complete linkage: maximum distances between all observations of the two sets
  - Average linkage: average of the distances of each observation of the two sets
  - Centroid linkage: distance between cluster centroids
  - Ward linkage: variance of the clusters being merged

HIERARCHICAL CLUSTERING

Python implementation

- Training a Hierarchical Clustering:

```
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean',
   linkage = 'average')
hc.fit(X)
```

- Predicting cluster attribution:

```
y_pred = hc.predict(X)
```

- Plotting the dendogram:

```
from scipy.cluster import hierarchy
dendrogram = hierarchy.dendrogram(hierarchy.linkage(X, method = 'ward'))
```

K-MEANS & HIERARCHICAL CLUSTERING

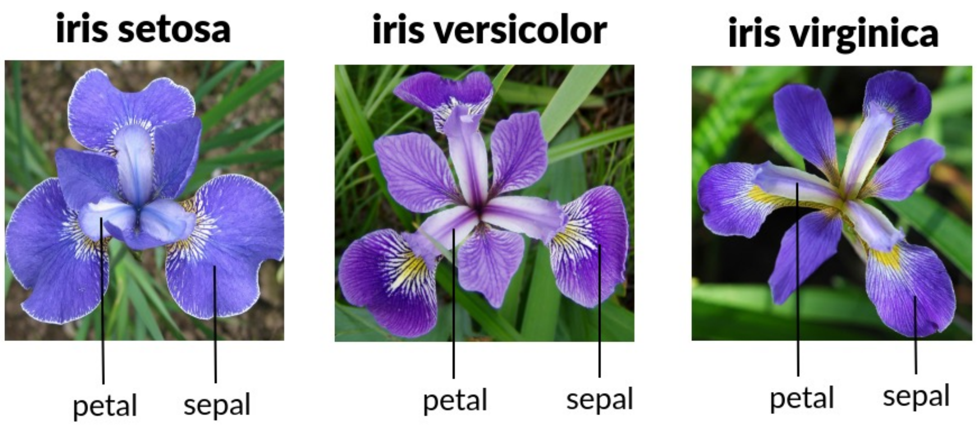
Implementation

- Data set: mall customer information
- Objectives:
  - Use unsupervised techniques to find segments of customers
  - Check visually the results

K-MEANS & HIERARCHICAL CLUSTERING

Student practice

- Data set: Iris data set: characteristics of three species of iris.
- Objectives:
  - Train a k-Means and Hierarchical Clustering
  - Visualize the results



- Check all previous lectures and practices and list your questions.