# Lecture 8: Unsupervised learning

## Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2023-2024

# Definition: unsupervised learning



## Question

Can anyone remind me what is the definition of **unsupervised learning ?**

# Definition: unsupervised learning



## Question

Can anyone remind me what is the definition of **unsupervised learning ?**

## Unsupervised learning

Unsupervised learning* is a type of algorithm that **learns patterns from untagged data**: through likeliness, algorithms build a concise representation of the data to generate imaginative content.

# Definition: unsupervised learning



While the goal of supervised learning is to **predict actions for unseen data**, the goal of unsupervised learning is to help us understand better **existing data**.

# Definition: unsupervised learning



While the goal of supervised learning is to **predict actions for unseen data**, the goal of unsupervised learning is to help us understand better **existing data**.

Several types of unsupervised learning exist in the literature:

- **Clustering methods**
- Latent models
- Anomaly detection

# Definition: unsupervised learning

While the goal of supervised learning is to **predict actions for unseen data**, the goal of unsupervised learning is to help us understand better **existing data**.

Several types of unsupervised learning exist in the literature:

- **Clustering methods**
- Latent models
- Anomaly detection

In this course, we'll focus on **clustering methods**.

# Definition: clustering

## Clustering

**Clustering** consists in **grouping a set of objects** so that objects in the same group (called a **cluster**) are more "similar" to each other than to those in other groups.
There is no class to be predicted but **the instances are to be divided into natural groups**.

# Definition: clustering



### Clustering

**Clustering** consists in **grouping a set of objects** so that objects in the same group (called a **cluster**) are more "similar" to each other than to those in other groups.
There is no class to be predicted but **the instances are to be divided into natural groups**.

Given a set of $j$ individuals described by their features $(x_{j,1}, \ldots, x_{j,n})$, assign each individual into a cluster $i$ $(1 \leq i \leq m)$.

# Definition: clustering



Clustering consists in finding **groups** where:

- Individuals **within** the group are similar
- Individuals **across** groups are dissimilar

# Definition: clustering



Clustering consists in finding **groups** where:

- Individuals **within** the group are similar
- Individuals **across** groups are dissimilar

There are many different clustering algorithms, that have a different interpretetation of "similar" and "cluster".

# Definition: clustering

Clustering consists in finding **groups** where:

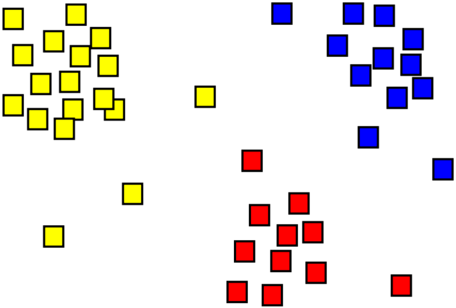
- Individuals **within** the group are similar
- Individuals **across** groups are dissimilar

There are many different clustering algorithms, that have a different interpretetation of "similar" and "cluster".

Similar to supervised learning, **there is no single best method for all datasets**.

# Example: clustering

| Height | Weight | Cluster |
| --- | --- | --- |
| 10 | 5 | ? |
| 8 | 3 | ? |
| 20 | 15 | ? |
| 17 | 16 | ? |



# Example use-cases



Possible use-cases for clustering:

- Finding groups within the data: streaming behavior, shopping behavior (market segmentation . . . ) . . .
- Finding outlier individuals in the dataset (individuals too far apart need to be investigated further)
- Semi-supervised learning: mapping samples to a set of class and using it for training.

# Definition: clustering



Selecting the right number of clusters can be done:

- Natively by the selected algorithm
- Iteratively by testing different values, evaluating different number of clusters and selecting the best

Clustering is also sensitive to overfitting: the variance-bias trade-off also applies here.

# Definition: clustering



Clustering can be:

- **Hard* clustering**: each individual belongs to **a single cluster**.
- **Soft* clustering** (*fuzzy*): an individual can belong to several clusters at the same time.

# Clustering algorithms



Any algorithms grouping a set of individuals into **groups of data** is a **clustering algorithms**.

# Clustering algorithms



Any algorithms grouping a set of individuals into **groups of data** is a **clustering algorithms**.

There are diverse algorithms:

- Centroid model based (k-means . . . )
- Connectivity models (hierarchical clustering . . . )
- Distribution-based clustering (latent models and gaussian mixtures . . . )
- Density based (DBSCAN . . . )

# Evaluation of clustering algorithms



**Question**

Do you think we can apply supervised learning metrics to the case of unsupervised learning ?

# Evaluation of clustering algorithms



**Question**

Do you think we can apply supervised learning metrics to the case of unsupervised learning ?

As we have no **ground truth**, we cannot directly use supervised learning metrics.

# Evaluation of clustering algorithms

**Question**

Do you think we can apply supervised learning metrics to the case of unsupervised learning ?

As we have no **ground truth**, we cannot directly use supervised learning metrics.
We need some metrics specific to the **unsupervised learning task**.

# Evaluation metrics

## Question

What could be some good metrics to assess the performance of an algorithm for a clustering task ?

Different possible approaches:

- Internal evaluation (**find a score describing the performance of the algorithm**)

# Evaluation metrics



## Question

What could be some good metrics to assess the performance of an algorithm for a clustering task ?

Different possible approaches:

- Internal evaluation (**find a score describing the performance of the algorithm**)
- Manual evaluation (**use a human expert to validate clusters meaning and see if they are consistent**)

# Evaluation metrics



## Question

What could be some good metrics to assess the performance of an algorithm for a clustering task ?

Different possible approaches:

- Internal evaluation (**find a score describing the performance of the algorithm**)
- Manual evaluation (**use a human expert to validate clusters meaning and see if they are consistent**)
- Empirical/indirect evaluation (**see if in practice the results of this clustering yields efficient information**)

# Clustering scores



Clustering scores assign the best score to the algorithm that produces clusters with **high similarity within a cluster** and **low similarity between clusters**.

Possible scores include:

- Silhouette scores

# Clustering scores



Clustering scores assign the best score to the algorithm that produces clusters with **high similarity within a cluster** and **low similarity between clusters**.

Possible scores include:

- Silhouette scores
- Davies-Boulin index

# Silhouette scores



## Silhouette

The silhouette score* measures **how similar an object is to its own cluster** compared to other clusters:
With $a$ the mean intra-cluster distance and $b$ the mean nearest-cluster distance, the silhouette score for an instance is:

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

# Silhouette scores

## Silhouette

The silhouette score* measures **how similar an object is to its own cluster** compared to other clusters:
With $a$ the mean intra-cluster distance and $b$ the mean nearest-cluster distance, the silhouette score for an instance is:

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

It ranges from -1 to 1: a high value indicates that the objects is **well matched to its own cluster** and **poorly matched to neighboring clusters**.

# Silhouette scores



If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value ($=$ individuals might have been clustered in the wrong cluster), then:

# Silhouette scores

If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value ($=$ individuals might have been clustered in the wrong cluster), then:

- Dataset may not be adequate for clustering

# Silhouette scores

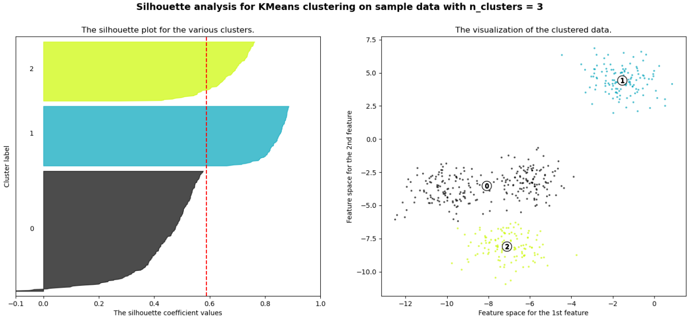
If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value (= individuals might have been clustered in the wrong cluster), then:

- Dataset may not be adequate for clustering
- Number of clusters may be poorly chosen (we will see for the different algorithms how to select the optimal number of clusters)

# Silhouette scores



Silhouette scores are usually visually represented as a *silhouette* plot to visually see how well the algorithm behaves.



# Silhouette scores: example



## Question

Given the following clustering result, compute silhouette scores (Manhattan distance) and plot the graph.

| ID | Height | Weight | Cluster |
|---|---|---|---|
| 1 | 10 | 16 | 1 |
| 2 | 12 | 14 | 1 |
| 3 | 14 | 15 | 1 |
| 4 | 14 | 30 | 2 |
| 5 | 30 | 30 | 2 |

# Davies-Bouldin index

## Davies-Bouldin index

The **Davies-Boulin index*** measures the average similarity of each cluster with its most similar cluster, where similarity is **the ratio of within-cluster distances to between-cluster distances**.

# Davies-Bouldin index



We define similarity between cluster $i$ and $j$ as:

$$R_{i,j} = \frac{s_i + s_j}{d_{i,j}}$$

with $s_i$ the average distance between each point of cluster $i$ and the centroid of that cluster and $d_{i,j}$ the distance between cluster centroids $i$ and $j$.

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

# Davies-Bouldin score: example

## Question

Given the following clustering result, compute Davies-Bouldin index (Euclidean distance) and plot the graph.

| ID | Height | Weight | Cluster |
|---|---|---|---|
| 1 | 10 | 16 | 1 |
| 2 | 12 | 14 | 1 |
| 3 | 14 | 15 | 1 |
| 4 | 14 | 30 | 2 |
| 5 | 30 | 30 | 2 |

# Intepretation of clusters





Usually as hard as the clustering task itself !

# Intepretation of clusters



Usually as hard as the clustering task itself !
Possible interpretations of clusters can be done:

# Intepretation of clusters



Usually as hard as the clustering task itself !

Possible interpretations of clusters can be done:

- By looking at the different scores to see how relevant clustering is

# Intepretation of clusters



Usually as hard as the clustering task itself !
Possible interpretations of clusters can be done:

- By looking at the different scores to see how relevant clustering is
- By looking at estimator values of the different features within each cluster

# Intepretation of clusters



Usually as hard as the clustering task itself !
Possible interpretations of clusters can be done:

- By looking at the different scores to see how relevant clustering is
- By looking at estimator values of the different features within each cluster
- By plotting the different clusters against the features and understanding why they were clustered together.

# Questions



Question ?