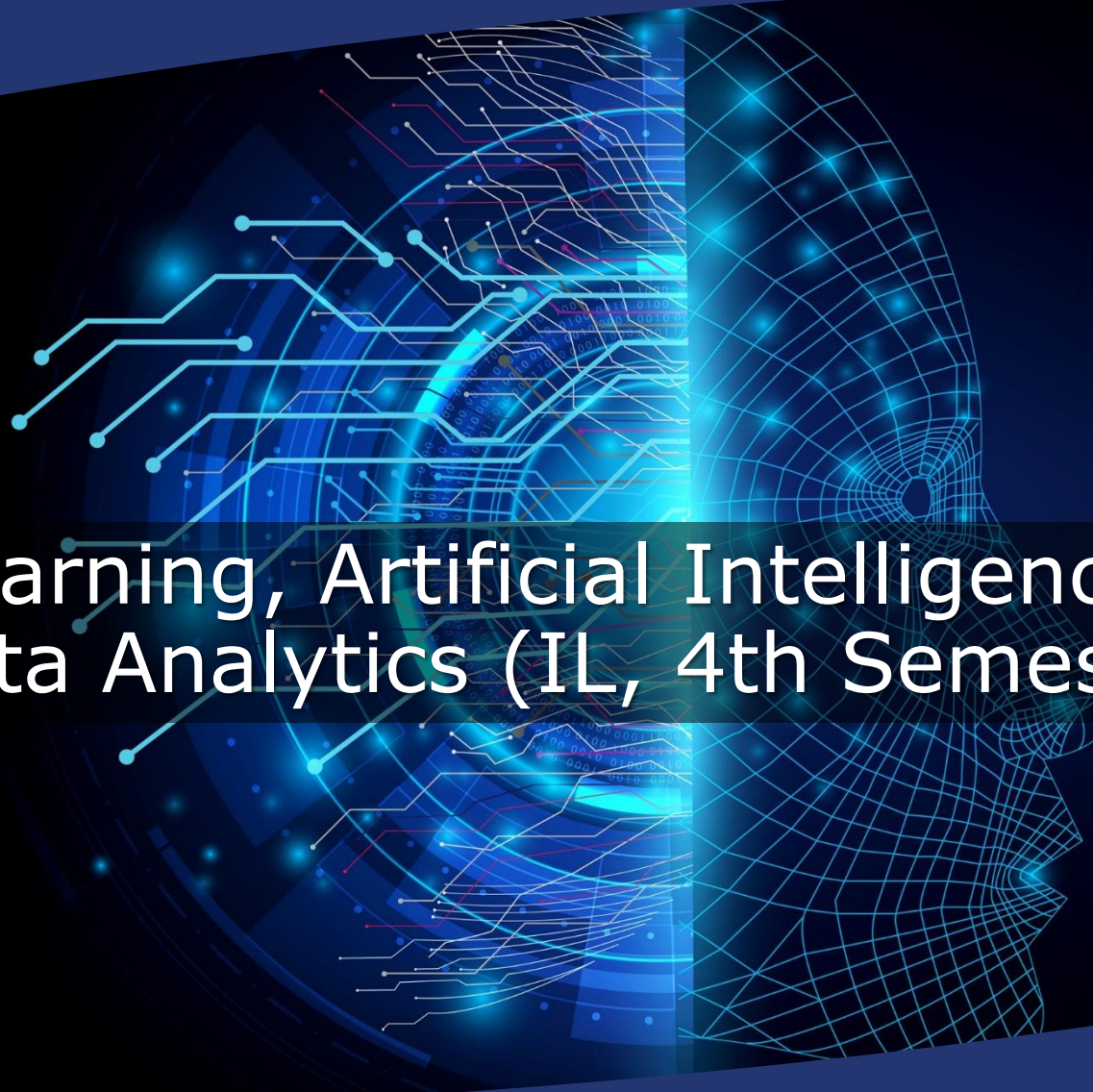
FH KREMS
UNIVERSITY OF APPLIED SCIENCES / AUSTRIA



Machine Learning, Artificial Intelligence, and Big Data Analytics (IL, 4th Semester)

Lecture 8

Where we are (Supervised learning)

- Regression and Classification, generalization error, data splitting, Bootstrapping, Bagging/Boosting, model validation, …
- K-Nearest Neighbors
- Support Vector Machines
- Decision Trees
- Random Forest
- Adaboost
- XGBoost

all supervised

structured data, tabular data = supervised learning

Coming next (Unsupervised learning)

- Introduction to unsupervised learning
- **Clustering** objective
- Partitioning methods
  - K-means
  - Partitioning Around Medoids (PAM)
- Hierarchical methods
  - Agglomerative (AGNES)
  - Divisive (DIANA)
- Density-based methods
  - DBscan and OPTICS

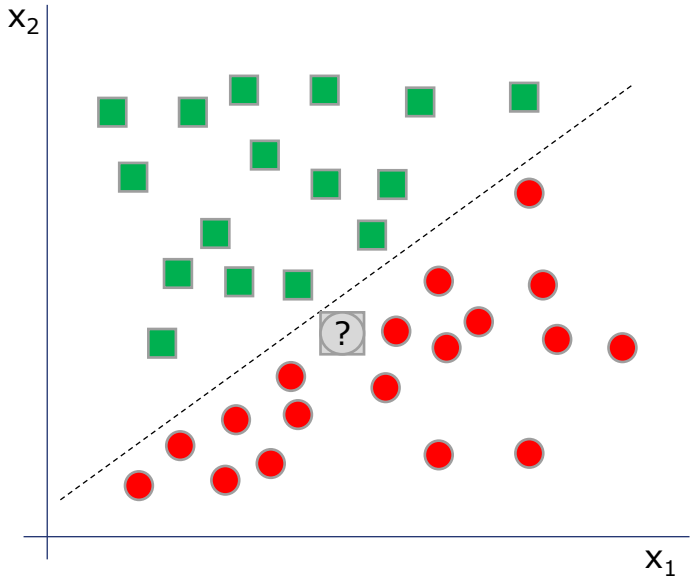
unstructered data = tries to find data without any relation

Agenda

- Introduction to unsupervised learning
- Clustering
- Types of Clustering
- K-means algorithm
- K-means variants
- Validating clustering

Supervised learning

Training data is labeled (e.g., green square vs. red circle)

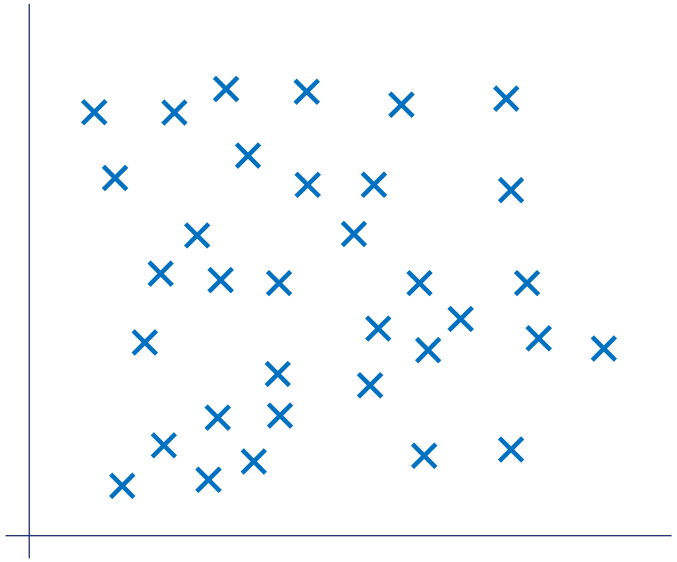


Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

GOAL: learn $f(x) \rightarrow y$

Unsupervised learning

Training data is not labeled



Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$

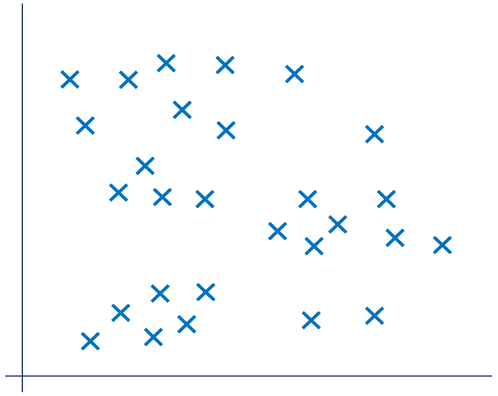
GOAL: find interesting things in data

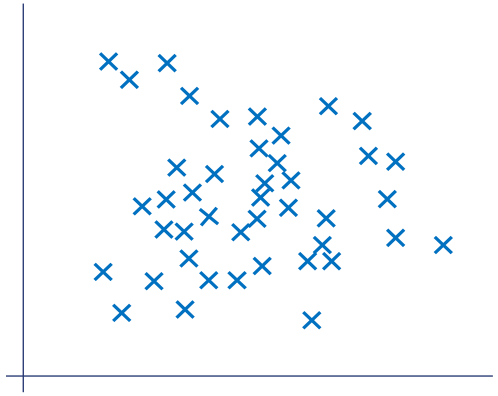
Examples of unsupervised learning techniques

- Clustering
- Dimensionality reduction
- Anomaly detection
- Association rules mining
- Pattern recognition

Clustering

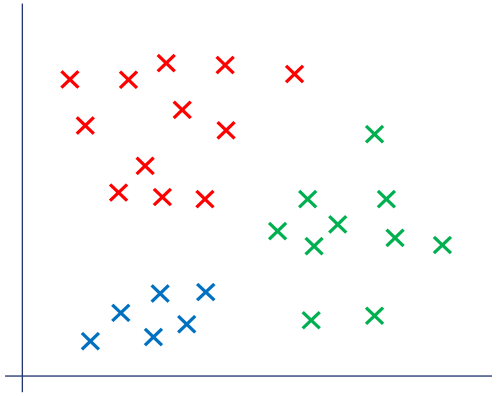
- Cluster analysis: Given a set of data objects find the proper grouping such that
  - Points in the same groups are similar to each other
  - Points in one group differ from points in other groups
- In other words: Finding natural groupings among objects in a dataset

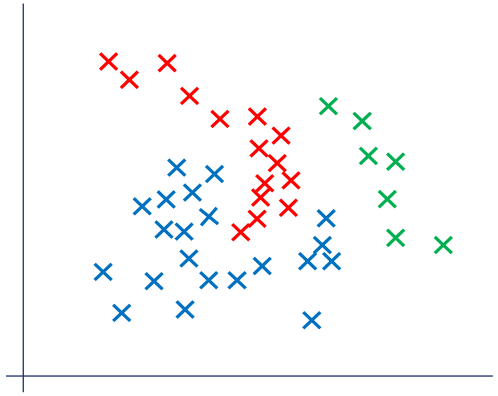




Clustering

- Cluster analysis: Given a set of data objects find the proper grouping such that
  - Points in the same groups are similar to each other
  - Points in one group differ from points in other groups
- In other words: Finding natural groupings among objects in a dataset

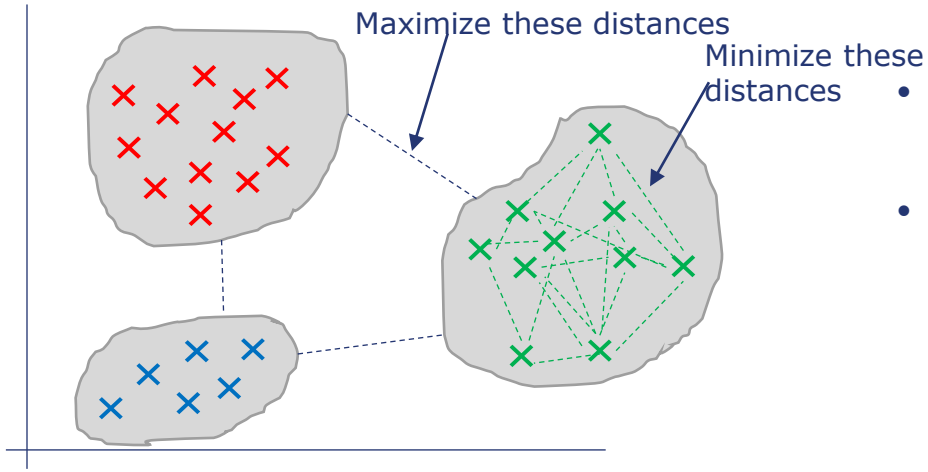




finding clusters of similiar points based on location

Clustering

- Optimal clusters should
  - Maximize similarity within clusters (intra-cluster): *cohesive* within clusters
  - Minimize similarity between clusters (inter-cluster): *distinctive* between cluster

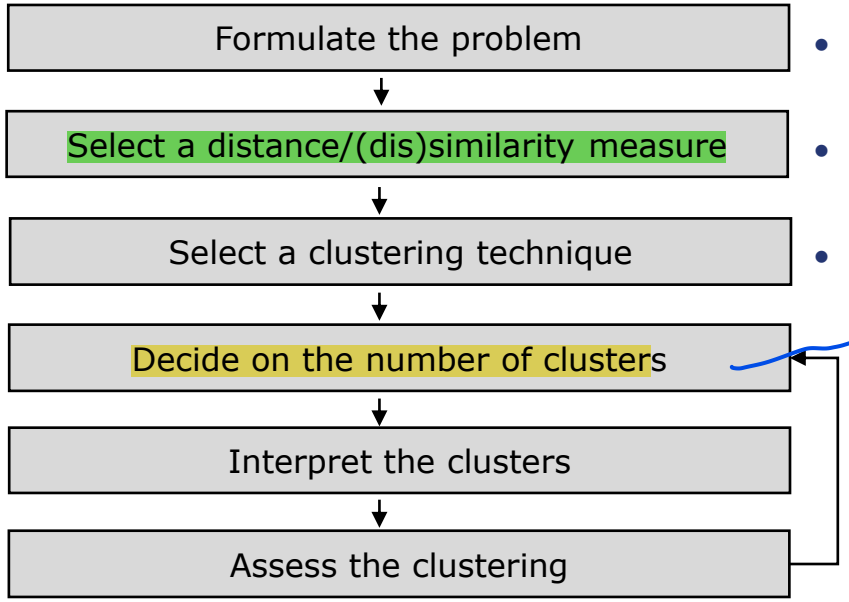


- dissimilarity is expressed in terms of a distance function, typically metric: $d(i, j)$
- The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables

Example of clustering

- Biology: Taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Behavioral: Understanding behavior of the masses
- Information retrieval: Document clustering
- Marketing: Discover distinct groups in their customer bases (customer segments), and use this knowledge to develop targeted marketing programs
- Climate: Understanding earth climate, find patterns
- Mobility: Understanding mobility patterns
- …

The clustering process



- Given a set of <…>, I want to find <…>, based on <…>
- Euclidean, Manhattan, Cosine, Jaccard, … last lecture
- K-means, PAM, Diana, DBScan, … model

overfitting underfitting

- Using Silhouette, Elbow, Dendograms, …

need to understand !

Types of clustering techniques

Types of algorithms

divide a dataset into non-overlapping subsets or clusters.
clusters are formed based on the similarity between the data points and the algorithm

- **Partitional algorithms:**
  - Construct various partitions and then evaluate them by some criterion.
  - Typical methods: k-means, k-medoids/PAM.
- **Hierarchical algorithms:**
  - Create a hierarchical decomposition of the set of objects using some criterion.
  - Typical methods: Diana, Agnes.
- **Density-based algorithms:**
  - Based on Connectivity and density functions
  - Typical methods: DBscan, OPTICS.
- …

Agglomerative, where each data point starts as its own cluster and is successively merged into larger clusters,
Divisive, where all data points start in
the same cluster and are successively split into smaller clusters.
The result is a tree-like structure that shows the relationships between the clusters at different levels of granularity.

looking for regions based on density

Types of clustering techniques

Hard vs. Soft Clustering

each data point is assigned to exactly one cluster, with no overlapping or sharing of clusters. In other words, hard clustering produces a crisp partition of the data.

- **Hard clustering**
  - Each sample belongs to exactly one cluster
  - For example: An *animal* belong to a *species*

- **Soft clustering**
  - A sample can belong to more than one cluster (probabilistic)
  - For example: *Glasses* belong to *medical aid* and to *fashion item*

Partitioning algorithms

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of intra-cluster squared distances is minimized
- Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
    - *k-means*: Each cluster is represented by the center of the cluster
    - *k-medoids* or PAM (Partition around medoids): Each cluster is represented by one of the objects in the cluster

K-Means algorithm

- Input:
  - K (number of clusters)
  - Training set $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$
- Algorithm
  - Randomly initialize K cluster centroids $\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{R}^n$
  - Repeat{

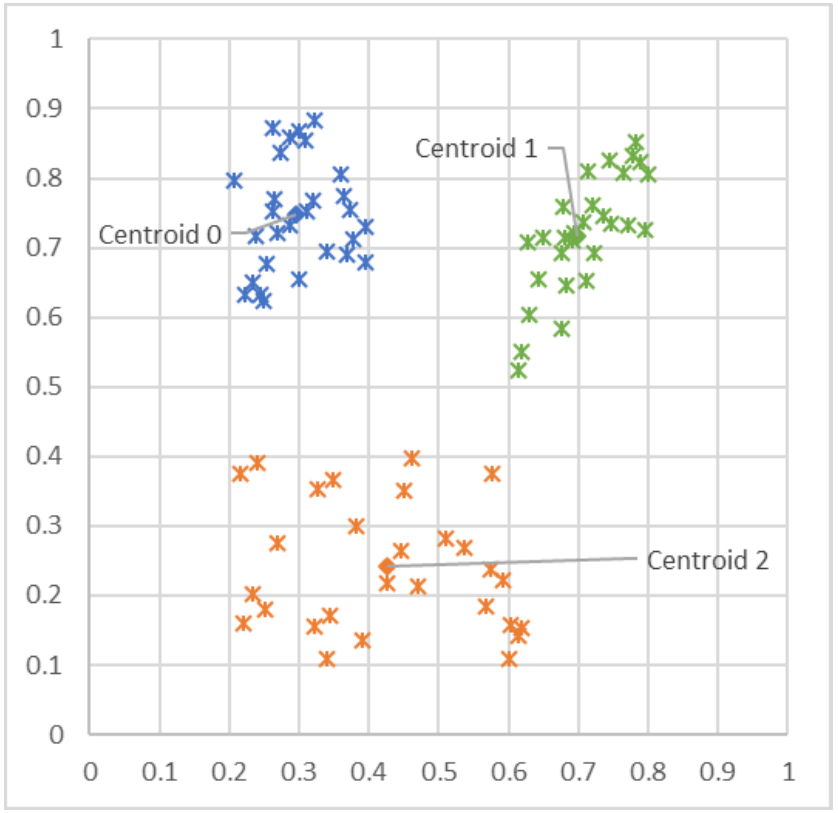
    for i = 1 to m — Cluster assignment step
    $c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

    for k = 1 to K — Centroid update step
    μ_k := average (mean) of points assigned to cluster k

  }

K-Means visual example

- Unlabeled Data
- Initialize centroids to random data point locations
- Each point assigned to nearest centroid
- Update centroid locations to avg of assigned points



- Each point assigned to nearest centroid
- Update centroid locations to avg of assigned points
- Each point assigned to nearest centroid
- Converge!

while loop

K-Means - How to choose K

- Pick a K based on your understanding of the domain
- Run K-Means
- Examine samples from each cluster
- Adapt K based on what you find
  - If single clusters contain different entities, increase K
  - If entities spread across clusters, decrease K
- OR
- Try multiple values of K and pick the K that maximizes a specified metric

K-Means considerations

- The objective of k-means is to minimize the total sum of the squared distance of every point to its corresponding cluster centroid.
- Finding the global optimum is NP-hard.
- The k-means algorithm converges to a **local** optimum.
- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.

K-Means weaknesses

- Sensitive to outliers
- Needs an initial guess on the number of clusters
- Results depend on the initial seeds
- Can be applied only to objects in a continuous n-dimensional space

Variants of K-means

- Many variants, usually differ by
  - Selection of the initial K
  - Initialization
  - Dissimilarity calculation
- Popular variants
  - K-medians:
    - Cluster center is the median
  - K-medoid:
    - Cluster center is an actual datapoint (not same as k-medians)
- Algorithms are similar to k-means

"doesnt use artificial point but when it updates centroid, it choses the point that is closest to the mean. "

chosing single point closest to the mean

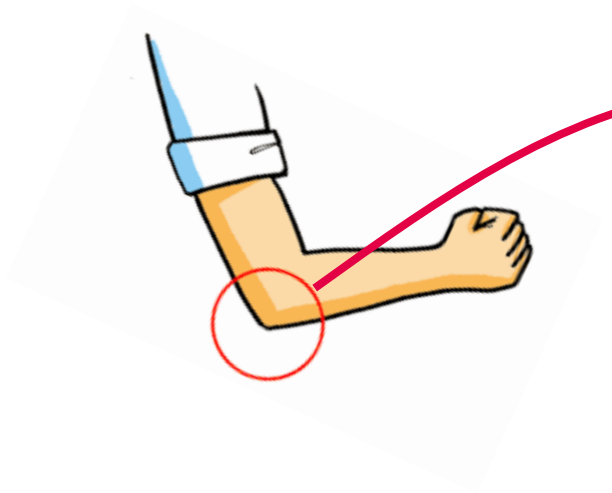
Evaluating the quality of clustering

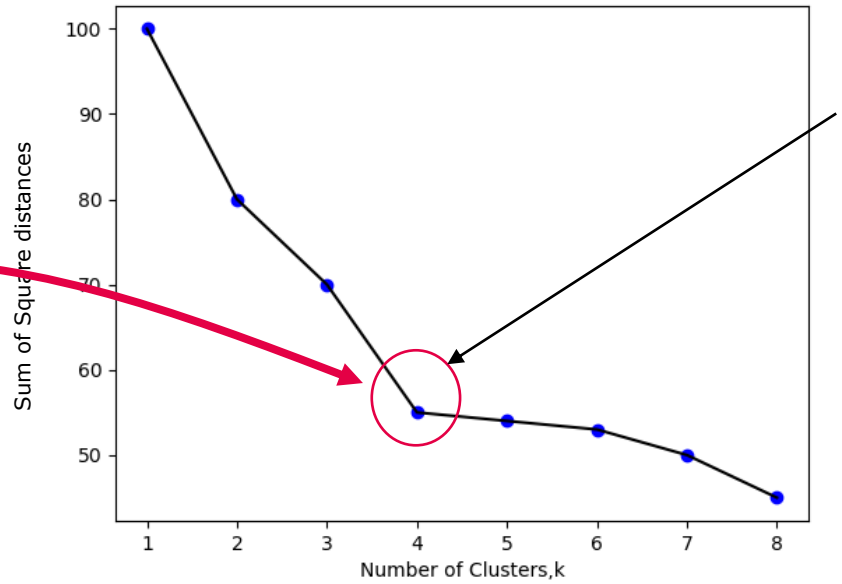
- Two type of measures: intrinsic and extrinsic
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are (e.g., intracluster similarity, intercluster dissimilarity, the Silhouette coefficient)
- Extrinsic: supervised, i.e., the ground truth is available for at least a subset of data
  - Compare a clustering against the ground truth using certain clustering quality measure

Evaluating the quality of clustering

The "Elbow method"

- SSE (or WSS): Sum of Squared Distances between data points and their assigned cluster's centroid





Elbow → n=4

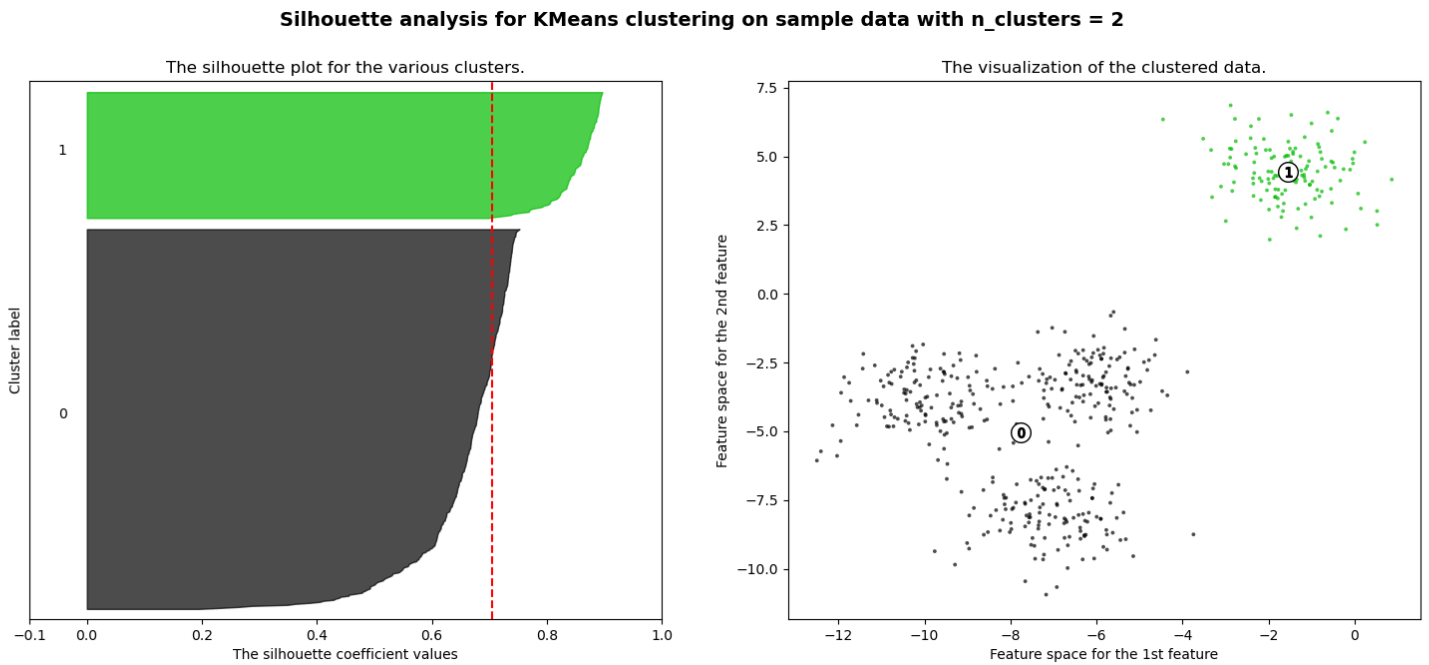
Evaluating the quality of clustering

The silhouette method

- Silhouette of one observation: $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$
- Where
  - *a(i) = average distance between i and all points in the same cluster*
  - *b(i) = average distance between i and all points in the closest different cluster*
- Interpretation:
  - *s(i) = 1* → observation fits well in its cluster and is far from other clusters
  - *s(i) = 0* → observation is as close to its cluster as to the neighbor cluster
  - *s(i) = -1* → observation is closer to the neighbor cluster than to its own
- Silhouette Score
  - SS=*mean{s(i)}* is a generic measure of cluster cohesiveness and distinction.

Evaluating the quality of clustering

The silhouette method (example)

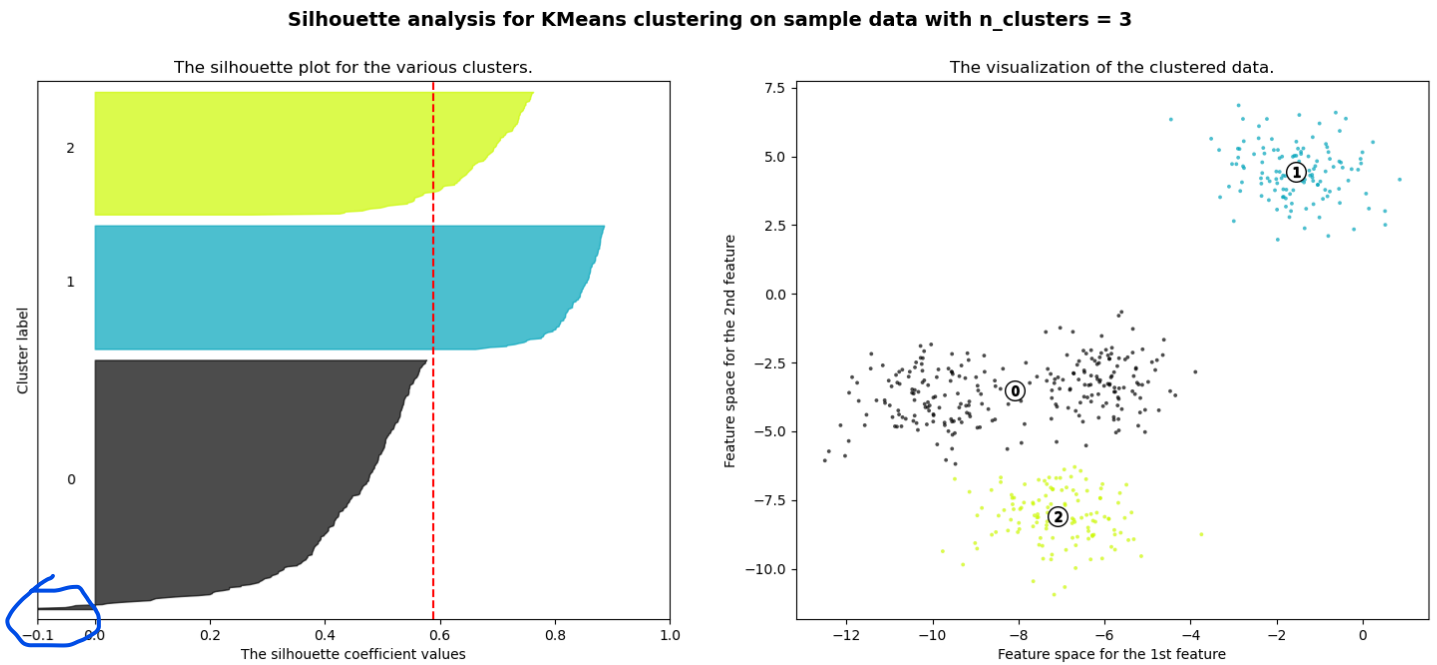


From: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Evaluating the quality of clustering

The silhouette method (example)

the longer the line the better the cluster

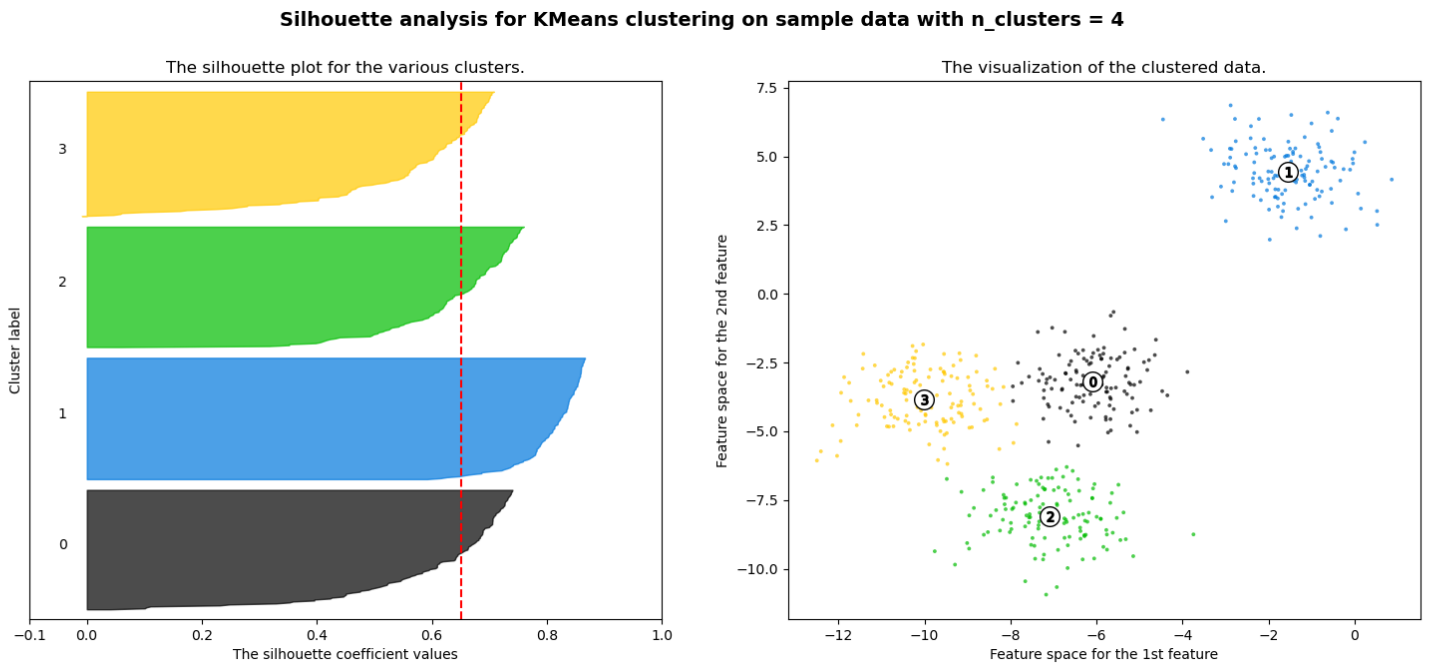


this is: we are closer to point in another cluster than the cluster we have selected

From: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Evaluating the quality of clustering

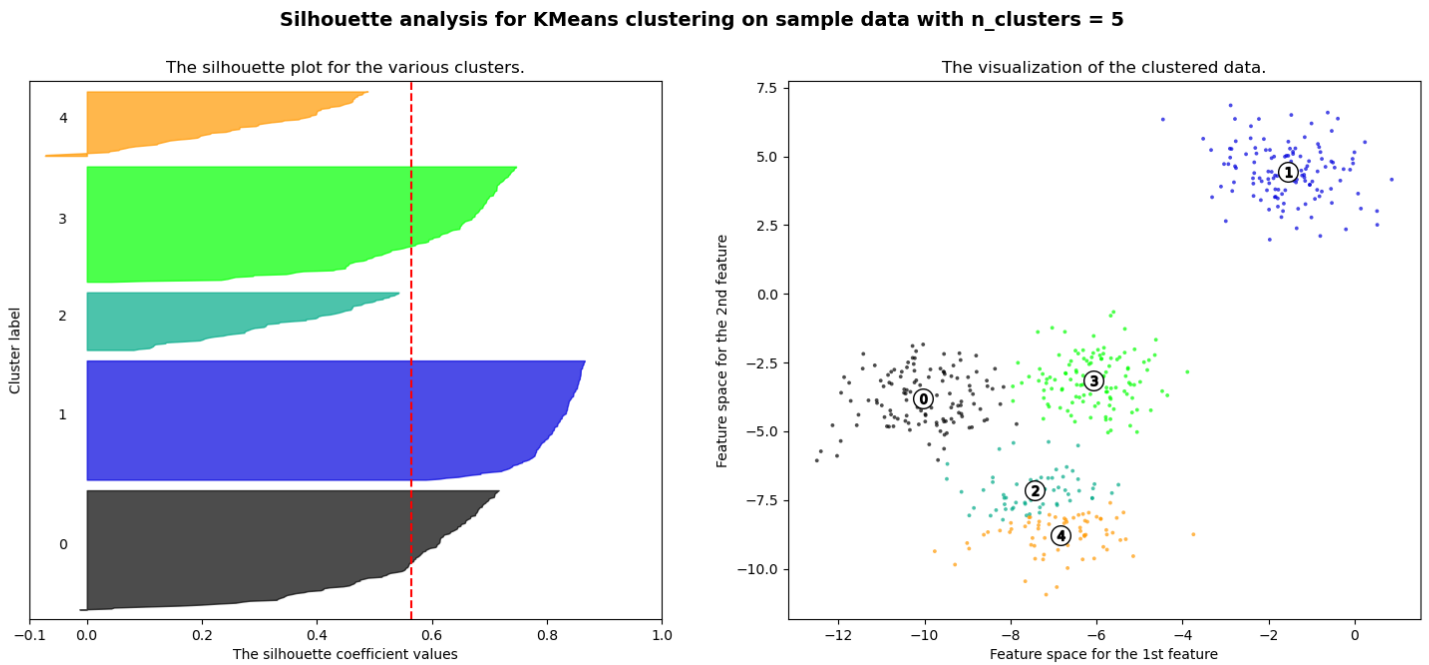
The silhouette method (example)



From: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Evaluating the quality of clustering

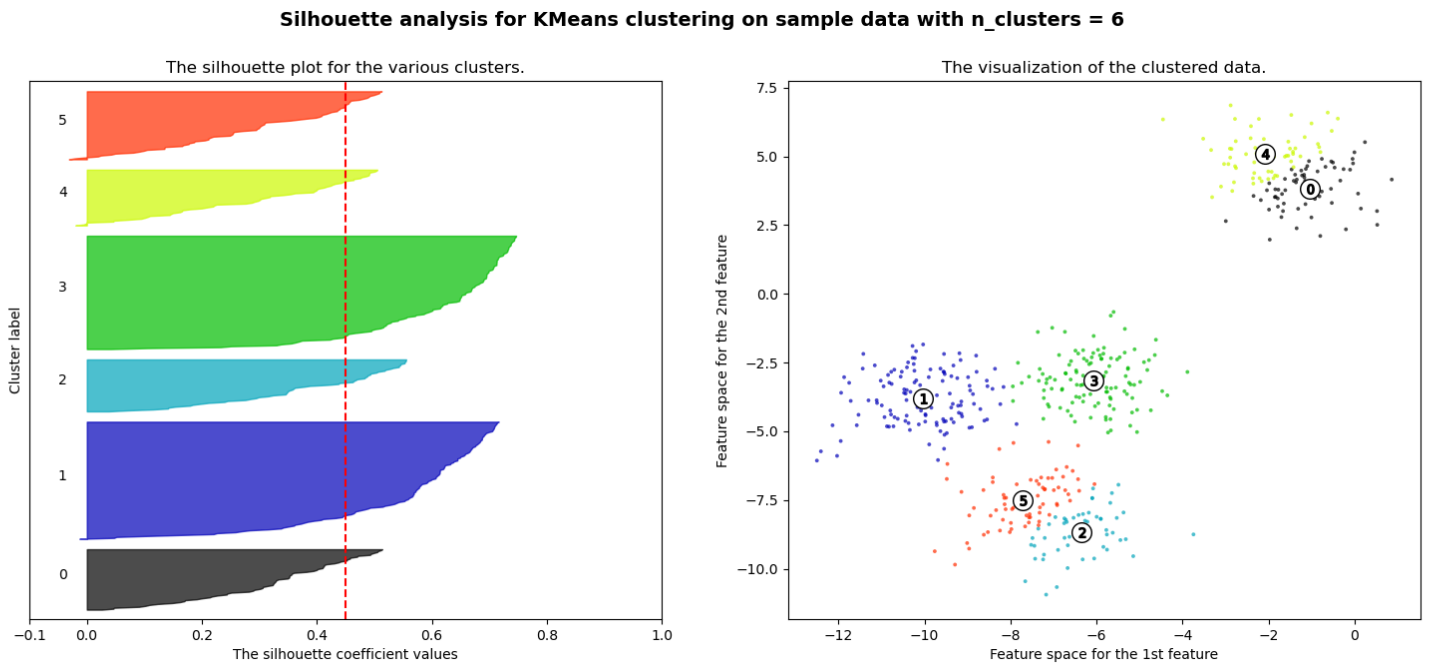
The silhouette method (example)



From: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Evaluating the quality of clustering

The silhouette method (example)



From: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Homework 3

DEADLINE: **May 8th, 09.00** (NO LATE SUBMISSIONS)
(MEGABONUS: **+5**) Do the assignment both in R and Python

- Create a report (**Rmarkdown and/or Jupyter notebook**) addressing the following assignments. The report must contain both code (cells) and results (no need to re-run)!

3.1 Mobile phone picture

- Take a picture with your smartphone. The picture must contain a piece of paper with your name on it and some type of background (walls, floor, window, etc.).
- Resize it to a manageable size (e.g., 256x256) either with R or Py
- The goal is to reduce the number of RGB colors by using k-means as in the lecture.
- Pick the k suggested by the elbow mechanism. Try also other k values.
- The report must contain the original pic, the WSS plot (elbow), & the final pictures

example

3.2 Drilling machine

- *drilling.csv* contains 400 operational measurements from a drilling machine.
- The machine can operate in different unknown states.

- Identify the number of states by using the known clustering techniques
  - K-means (iterate over k → elbow → final clustering)
  - Hierarchical clustering (iterate over linkages → AC/dendrograms → final clustering)
  - Dbscan (kNNdistplot → Eps → final clustering)
  - Optics
- The report must contain all plots and a final comparison of the different clustering outcomes

- Coding session