LECTURER: TAI LE QUY

MACHINE LEARNING

UNSUPERVISED LEARNING AND FEATURE ENGINEERING

UNIT 2

CLUSTERING

- Explain the **functioning principal of clustering** approaches and how they work.
- **Implement** a clustering approach.
- **Test and evaluate the quality** of the obtained clusters.
- **Choose the clustering approach** with respect to the challenges and constraints of the dataset.

EXPLAIN SIMPLY

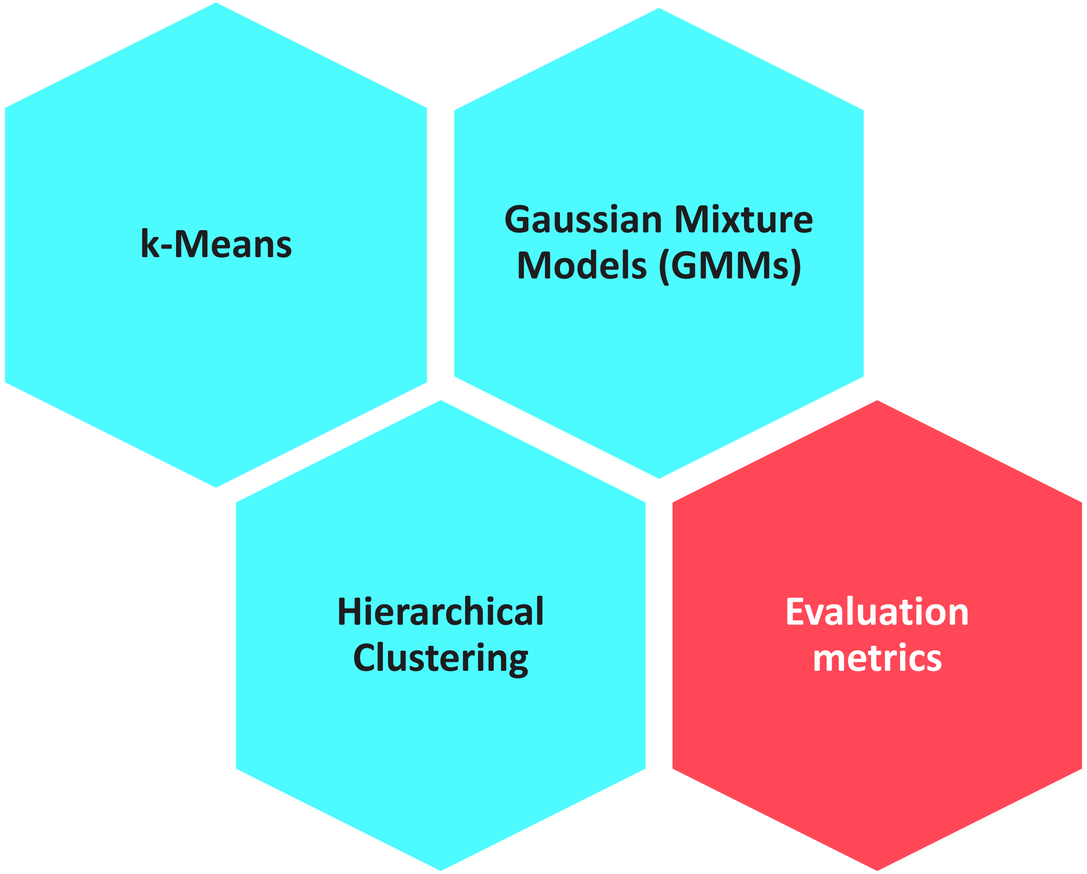
1. Explain how it is possible to obtain **two different clustering results** for the **same dataset** using **k-Means** clustering.

2. In k-Means, the **centroids** are **updated** in each iteration. Explain the equivalent in **Gaussian Mixture Models** that is updated in each iteration.

3. For a **100-sample dataset**, explain **how many samples** will be in **each leaf** and how many will be in the **stem** of the dendrogram when applying hierarchical clustering.

UNIT CONTENT

Image 1: Unit content - Clustering

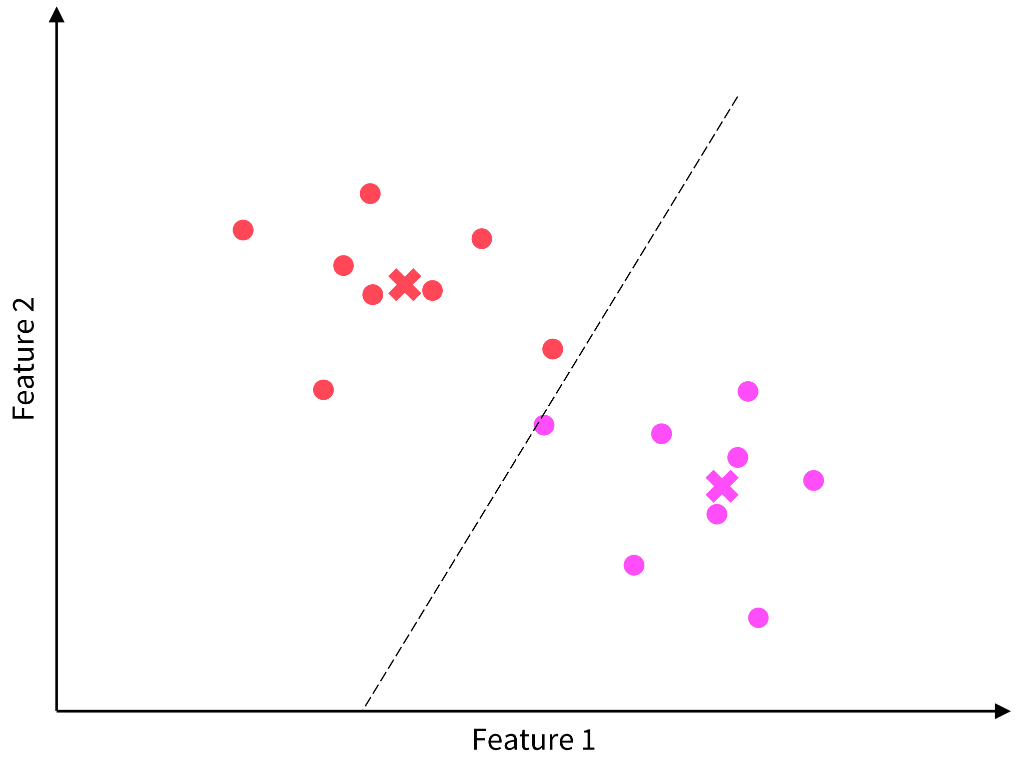


Source of the image: Müller-Kett, 2021.

K-MEANS

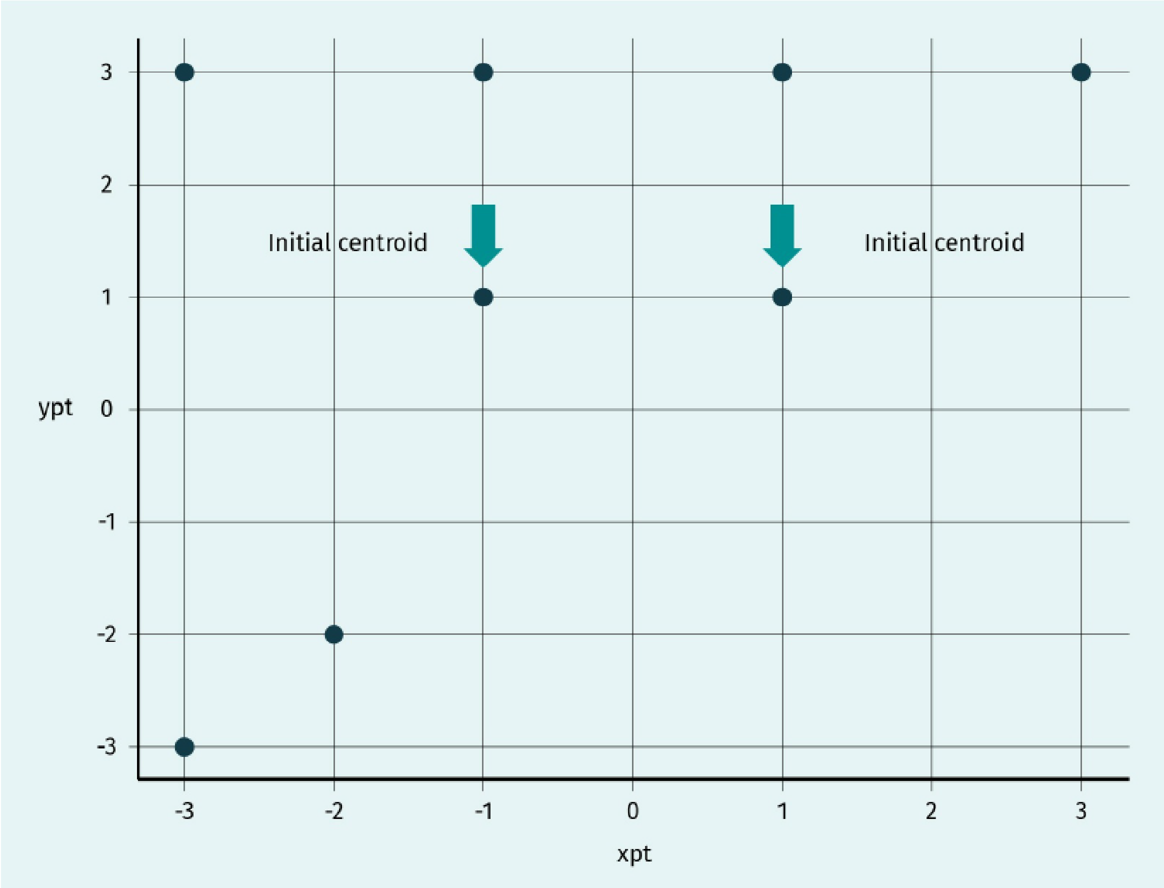
1. Choose a **number of clusters, k**
2. **Randomly select** a data point for each cluster (**seed = interim centroid**).
3. Calculate the **distance** between **each data point** and the **centroids.**
4. Assign each data point **to the nearest centroid.**
5. Select new centroids as the **mid-point** of each cluster.
6. Repeat steps 3 to 5 until the **stop criterion is fulfilled.**

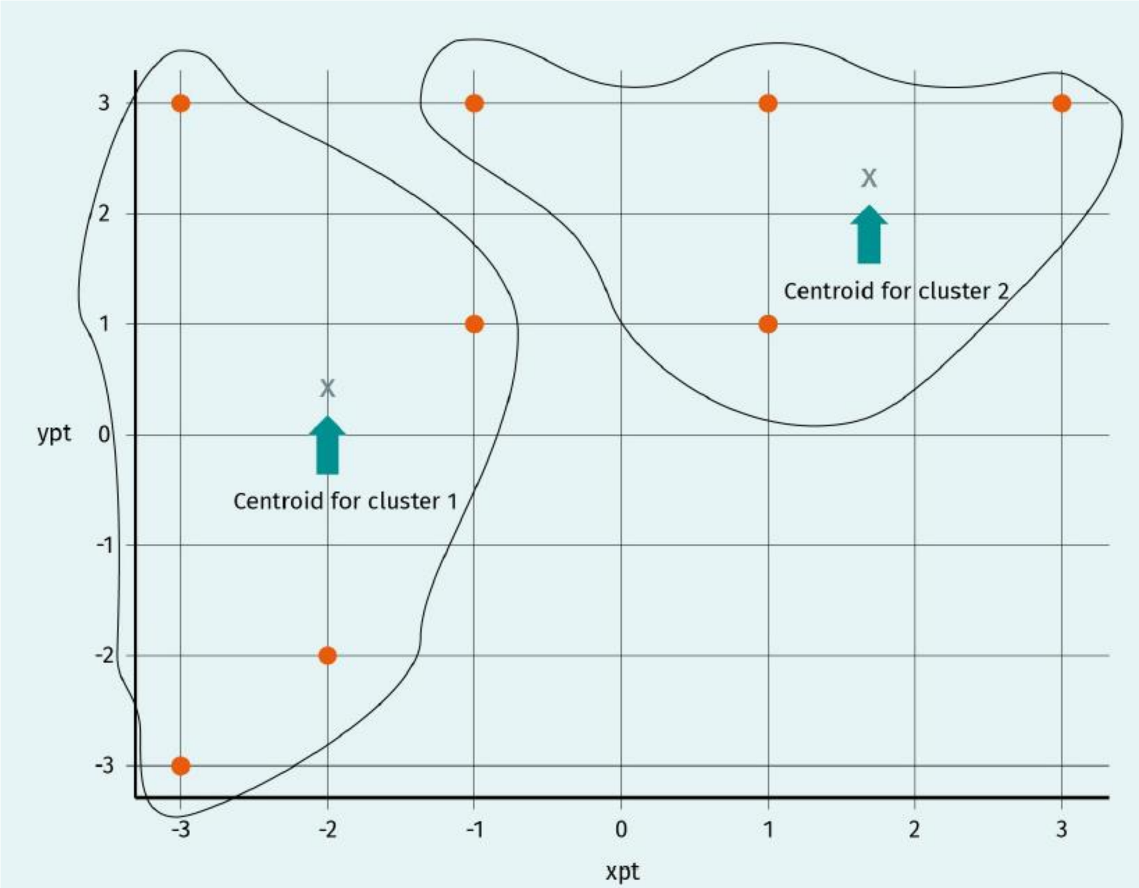
Image 2: k-Means



Source of the image: Müller-Kett, 2020.

K-MEANS - EXAMPLE





$$d\left(a,b\right) = sqrt\left[\left(a_x - b_x\right)^2 + \left(a_y - b_y\right)^2\right]$$

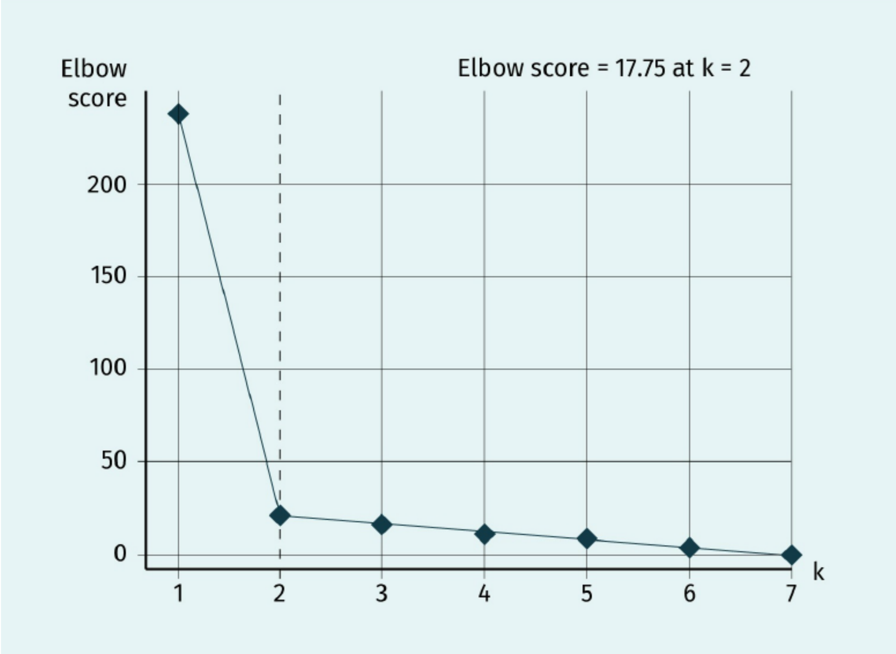
- k-Means is **not deterministic.**
- There are several **variations** to k-Means.
- for large datasets
  - **Clustering for Large Applications (CLARA)**
    - Partitioning
    - Batch-processing

Source of the image: Müller-Kett, 2020.

- The number of clusters to discover in the dataset is defined by the user.
- Methods:
  - Visualization
  - Domain knowledge
  - Data-driven approaches: use a metric to determine the quality of the obtained clusters for different values of k
    - Elbow method
    - Silhouette score

- Looks for the number of clusters ***k*** for which adding more clusters will not add considerable information to increase the quality of the clustering and give better modeling results with respect to the data variation.
- The value of k that gives the cutoff or "**elbow**" of the curve, when plotting the cost against ***k***, is chosen.
- Total within-cluster sum of squares (WSS)

$$WSS = \sum_{j=1}^{k} \sum\nolimits_{x_i \in c_j} \left(x_i - c_j\right)^2$$



Source of figure: course book

SILHOUETTE SCORE

- Mean distance a(i) between data point x_i and all other data points in the same cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} dist(x_i, x_j)$$

- Smallest mean distance b(i) between data point x_i and any other data point in any other cluster:

$$b(i) = \min_{j \neq i} \frac{1}{|C_j|} \sum_{x_j \in C_j} dist(x_j, x_j)$$

- Silhouette score s(i) for the data point x_i in cluster C_i

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, if a(i) < b(i) \\ 0, if a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, if a(i) > b(i) \end{cases}$$

s(i) close to 1: x_i is appropriately clustered

s(i) close to zero: x_i is on the border of two natural clusters

s(i) close to -1: x_i is badly clustered

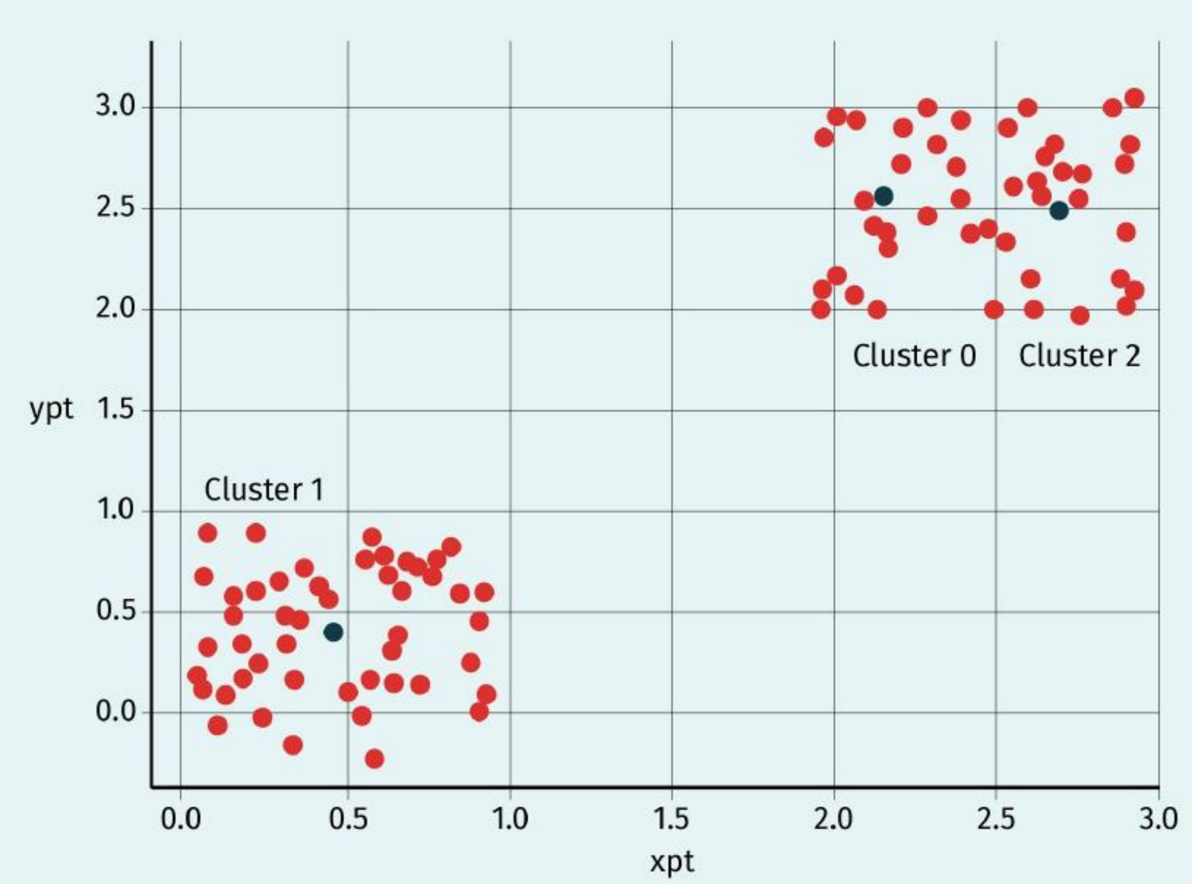
- Silhouette of a cluster $\overline{s}(C_i) = \frac{1}{|C_i|} \sum_{X_i \in C_i} s(X_i)$
- Silhouette of a clustering $s(k) = \max_{j = 1, \ldots, k} (\overline{s}(C_j))$

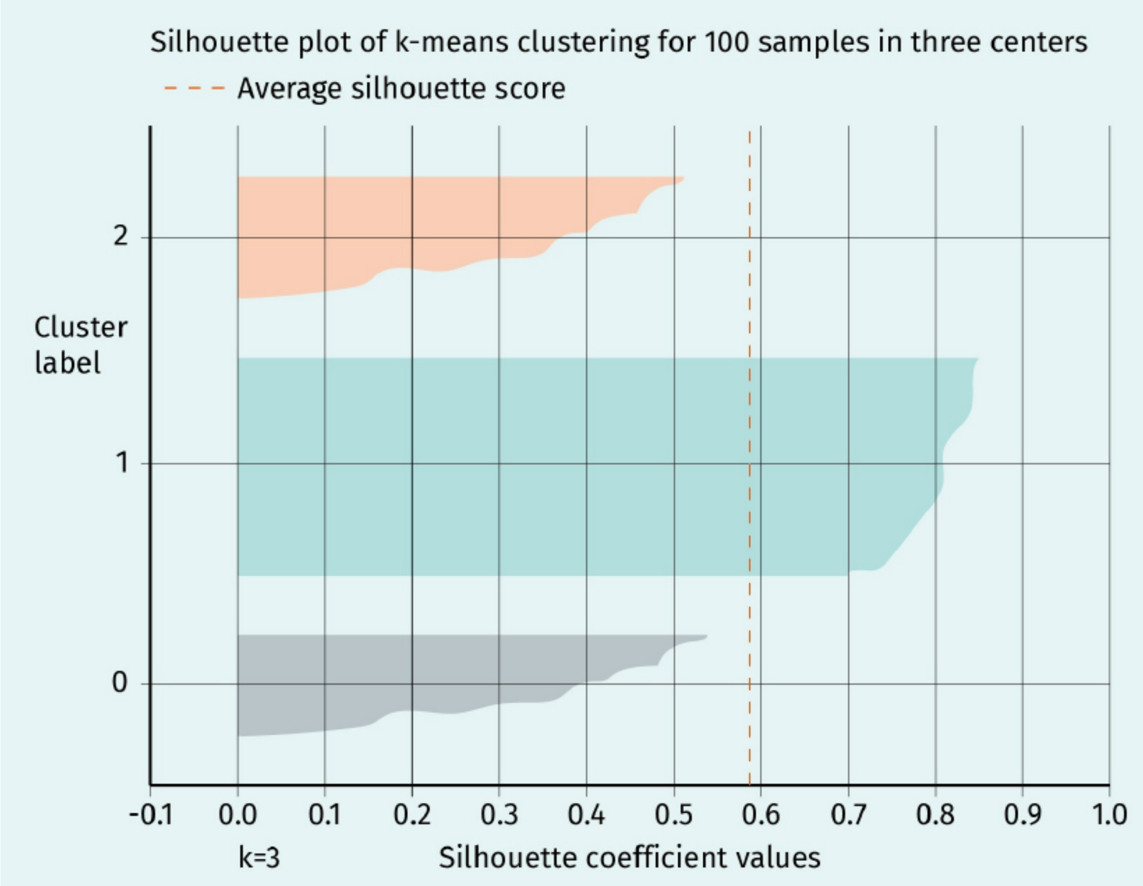
s(k) close to 1: data point are well clustered

s(k) close to zero: clusters are indifferent

s(k) close to -1: data points are badly clustered

SILHOUETTE SCORE

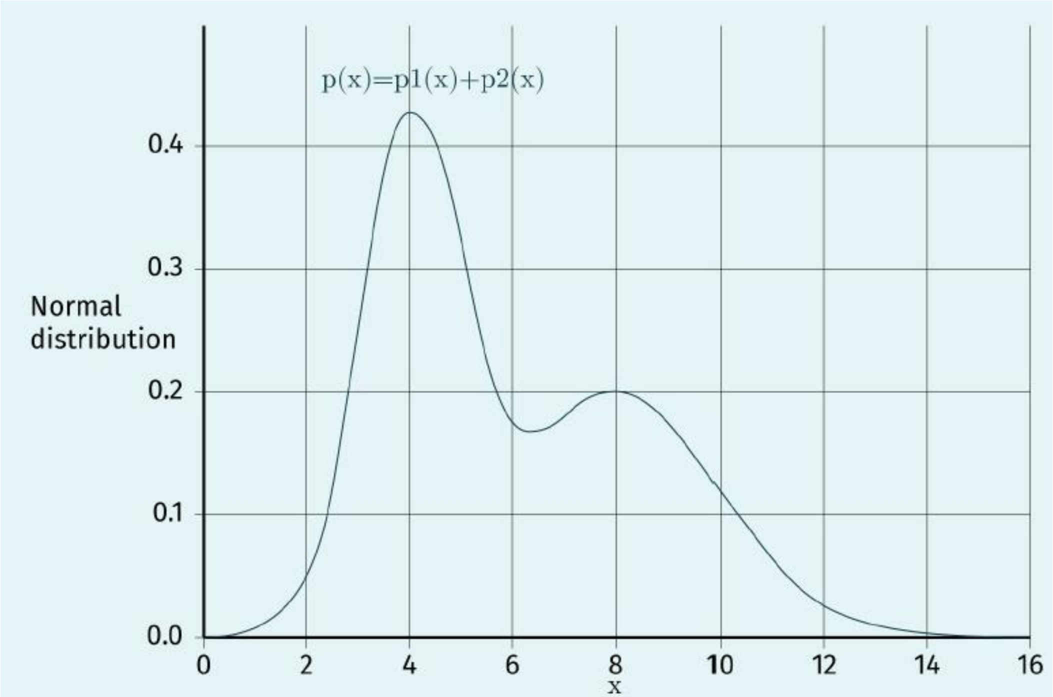


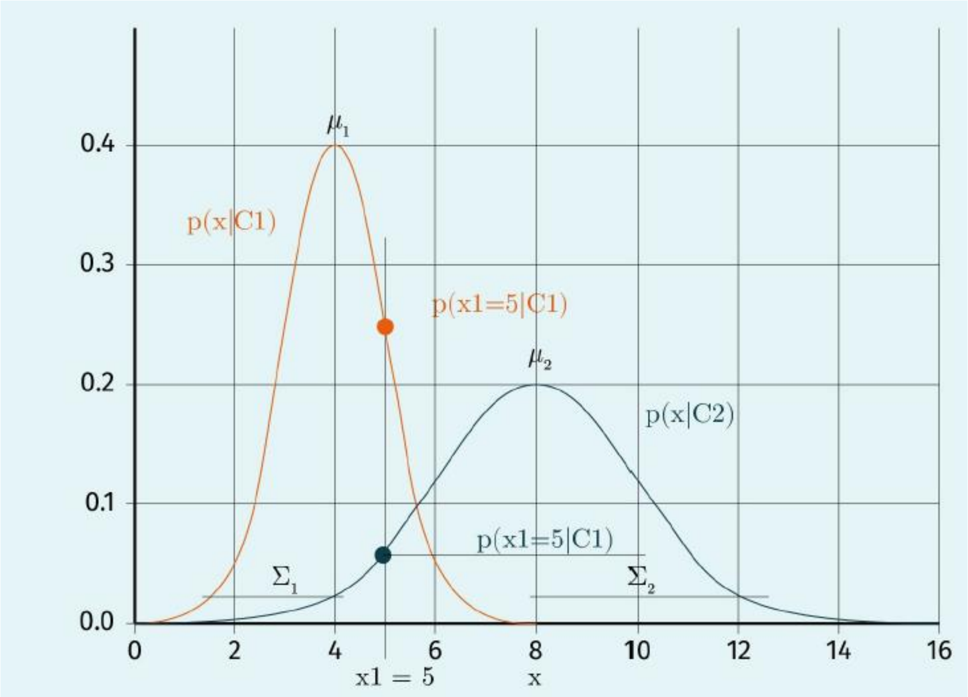


Two Gaussian Clusters and Corresponding Silhouette Measure

Source of figures: Course book

- A soft or probabilistic clustering method
- Each cluster is represented by a prototype (a Gaussian or normal probability density $p(x|Cj)$, $j = 1, \ldots, k$, represented by its two parameters: the mean value μ_j and the variance-covariance matrix)





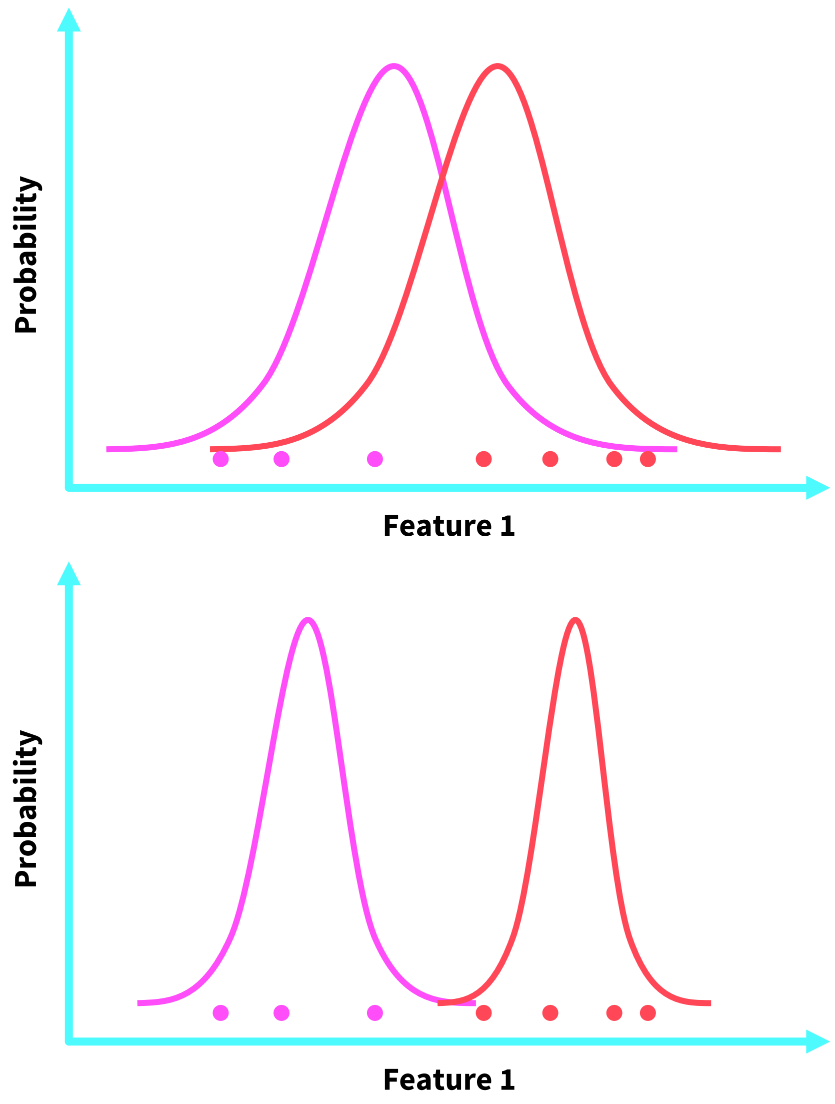
Two Clusters Represented by Two Gaussian Probability Densities and Their Mixture Probability Density

Source of figures: course book

GAUSSIAN MIXTURE MODEL CLUSTERING (GMM)

1. Choose „**prior probabilities**“ at random.
2. Assign **each sample** to the **closest cluster centroid** based on the **Maximum Likelihood.**
3. **Re-calculate** the cluster **centroids** based on the **mean and variance** of the samples in this cluster.
4. Repeat steps 2 and 3 until the **stop criterion is fulfilled.**

Image 3: GMM



Source of the image: Müller-Kett, 2020.

1. Represent each cluster C_j, j = 1, …, k, with a Gaussian or normal probability density $p(x|C_j)$, j = 1, …, k, and its prior probability π_j.
2. Define the mixture probability p(x) that a data point x belongs to k clusters.

$$p(x) = \sum_{j=1}^{k} \left(p\left(xC_j\right) . \pi_j\right)$$

3. Estimate the normal or Gaussian probability density parameters (μ_j, Σ_j, π_j) for each cluster C_j. (expectation-maximization (EM) algorithm)
   - **Expectation:**
     - The parameters (μ_j, Σ_j, π_j) of each cluster C_j are initialized. Compute the posterior probability of x

$$p\left(C_j x\right) = \frac{p\left(x \middle| C_j\right) . \pi_j}{\sum_{i=1}^{k} \left(p\left(x \middle| C_i\right) . \pi_i\right)}$$

     - Used to assign a data point x to a cluster C_j

- **Maximization**
  - The parameters (μ_j, Σ_j, π_j) of each cluster C_j, will be updated using the weighted data points by the posterior probabilities

$$\mu_j = \frac{\sum_{i=1}^{n} p(C_j | x_i) . x_i}{\sum_{i=1}^{n} p(C_j | x_i)}$$

$$\Sigma_j = \frac{1}{\sum_{i=1}^{n} p(C_j | x_i)} \sum_{i=1}^{n} p(C_j | x_i) . (x_i - \mu_j)^T . (x_i - \mu_j)$$

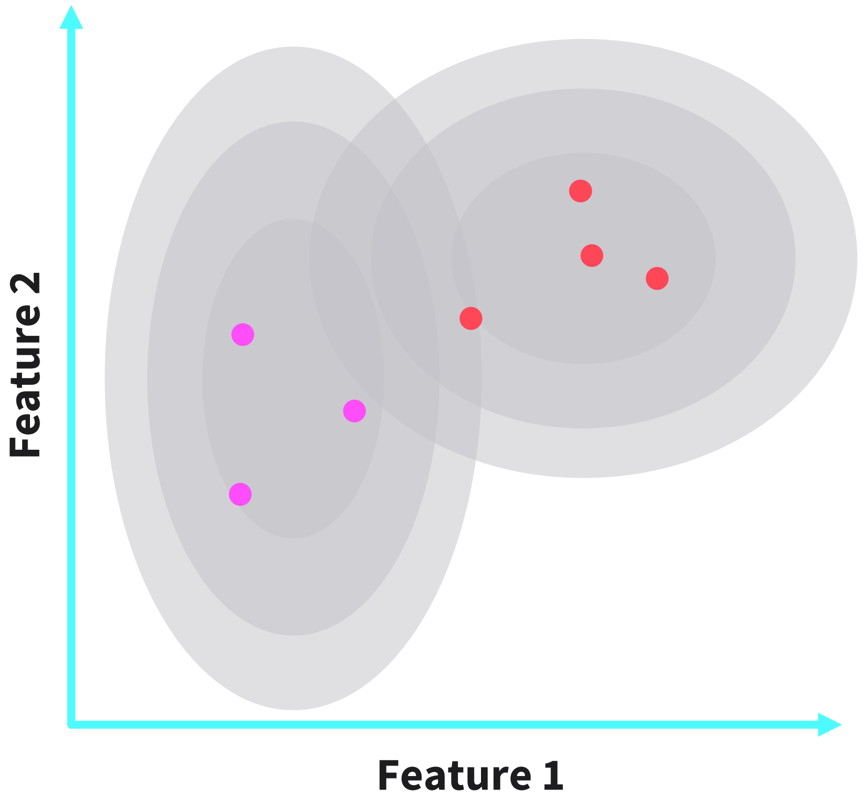
$$\pi_j = \frac{\sum_{i=1}^{n} p(C_j | x_i)}{n}$$

4. The process will restart with the expectation step until the EM converges to where no change occurs in the parameters

Advantages

- **"Fuzzy" clusters** with probability zones.
- Different **"probability slopes" for each feature.**

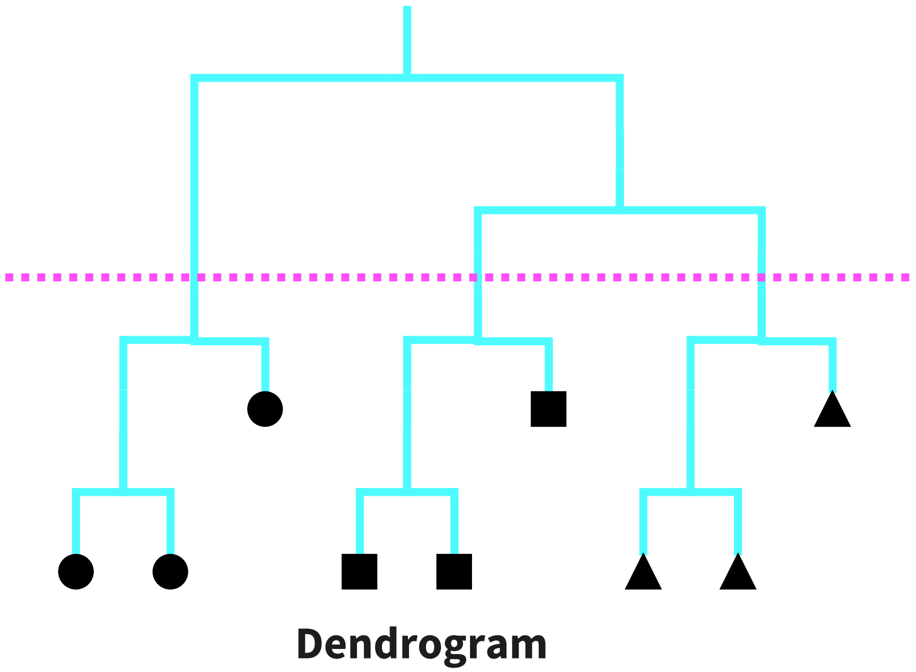
Image 4: GMM clusters with probability zones



Source of the image: Müller-Kett, 2020.

1. **Each sample** in one **cluster** (leaves).
2. Calculate **distances** between all samples.
3. Group the **two closest samples**, respectively.
4. Continue **grouping samples and groups.**
5. Ultimately, all sample end up in **one cluster** (stem).
6. Choose the **number of clusters** by horizontally drawing the **decision boundary** through the dendrogram.

Image 5: Hierarchical clustering

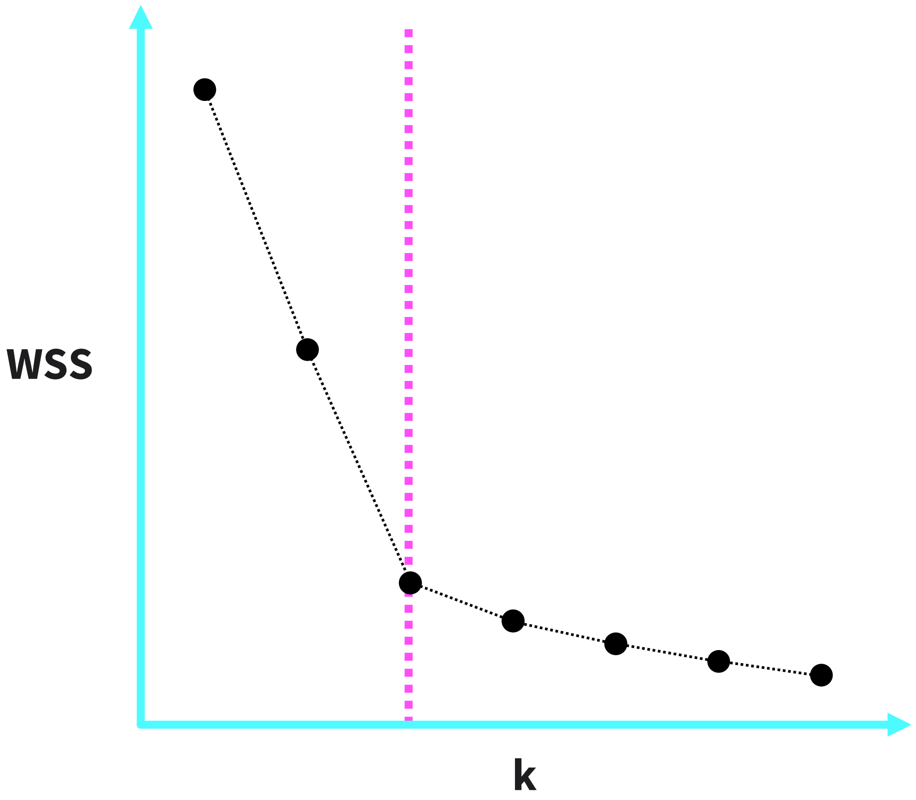


Source of the image: Müller-Kett, 2020.

Elbow method with WSS

- **Within-Cluster Sum of Squares (WSS)**
  - Squared **distance** between **data points** and respective cluster **centroids.**
  - $WSS = \sum_{j=1}^{k} \sum_{x_i \in c_j} (x_i - c_j)^2$

Image 6: Elbow method

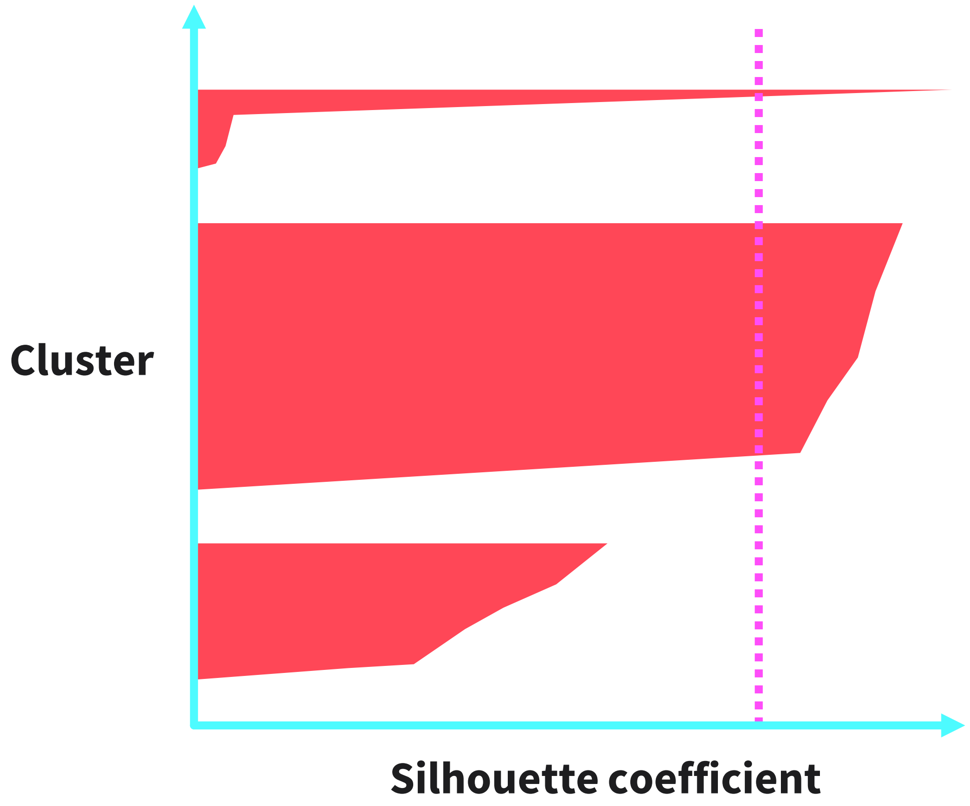


Source of the image: Müller-Kett, 2023.

Silhouette Score

- Cohesion
- Separation
- Range [-1, 1]
- For each data point
- As overall metric

Image 7: Silhouette score

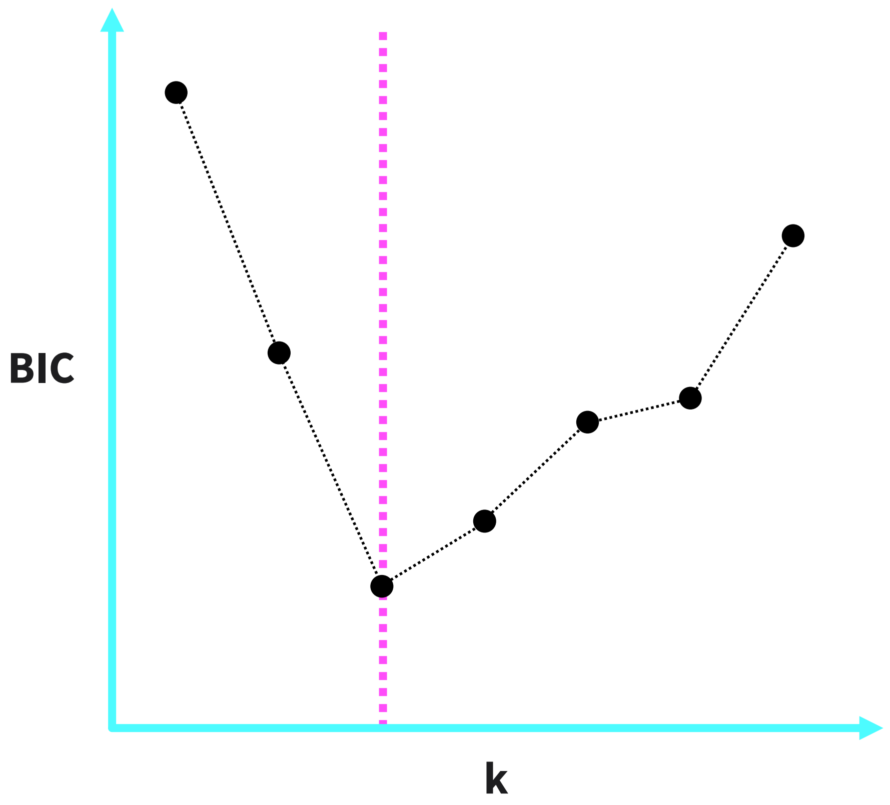


Source of the image: Müller-Kett, 2023.

Bayesian Information Criterion (BIC)

- $BIC = ln(n) \cdot p - 2ln(L)$
- n = number of samples
- p = number of parameters
- L = Maximum Likelihood

Image 8: BIC



Source of the image: Müller-Kett, 2023.

- Explain the **functioning principal of clustering** approaches and how they work.
- **Implement** a clustering approach.
- **Test and evaluate the quality** of the obtained clusters.
- **Choose the clustering approach** with respect to the challenges and constraints of the dataset.

SESSION 2

TRANSFER TASK

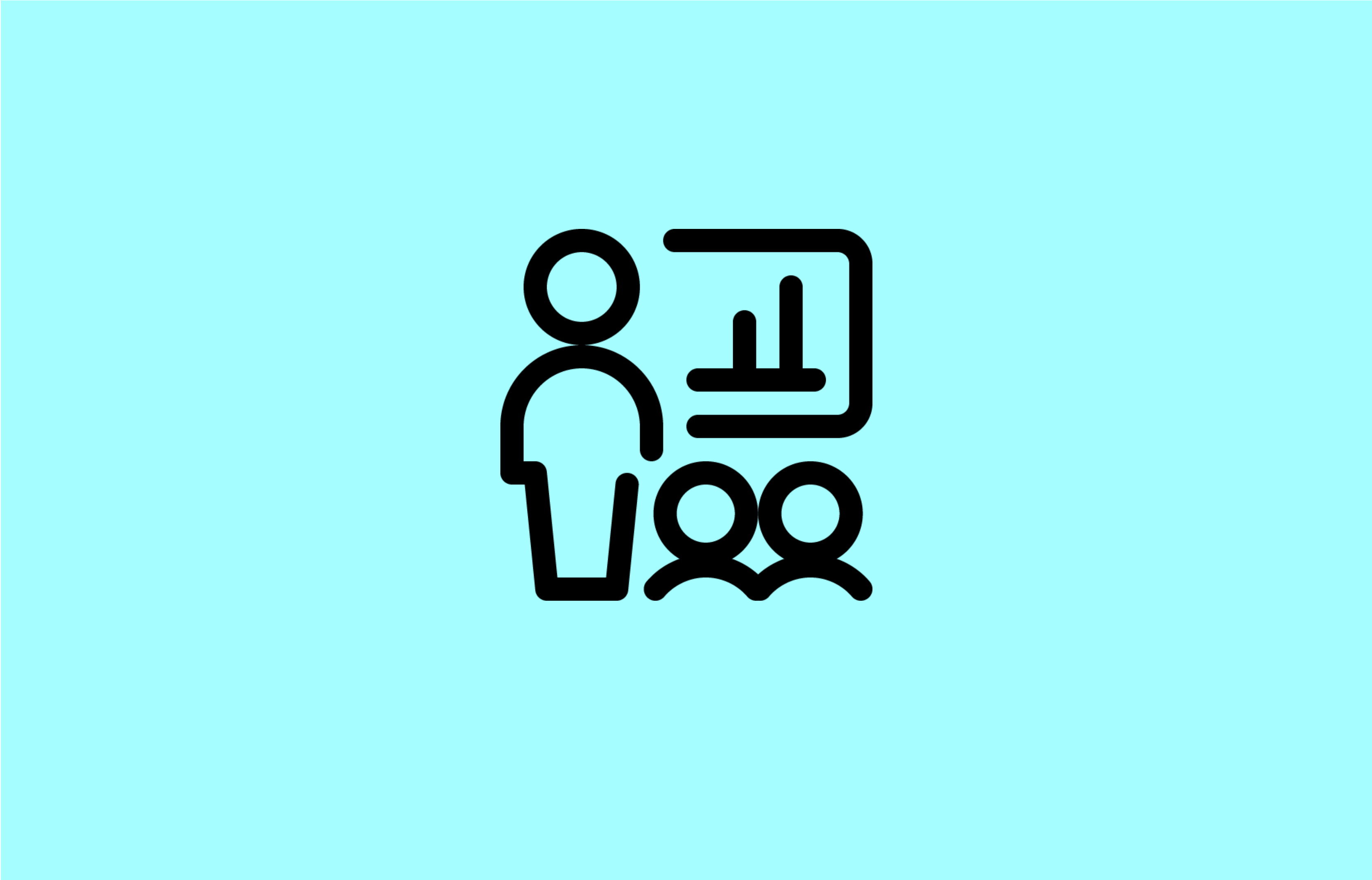
TRANSFER TASKS

A start-up that sells **sustainable products in smaller stores** has been very successful in recent years. As a result, more stores are to be opened worldwide.

To keep an **overview of the offered products**, you and your team of Data Scientists are tasked to **define homogeneous groups of products** to facilitate ordering, marketing, and distribution. There are **different use cases** for your results, and you should use different methods appropriate to each use case:

1. The customer base **constantly changes**, and the clustering must be conducted **as quickly as possible**.
2. Once a month, a **more thorough analysis** should be conducted. **Not all features** seem to be **equally informative** to differentiate the customers into groups.
3. The **number of clusters** has to be **adapted on-the-fly** for the ordering process to quickly assess how many different products should be ordered in bulk.

Please present your results.
The results will be discussed in plenary.



1. What does the elbow criterion consider when assessing the quality of clusters?
   a) the cohesion of the clusters
   b) the separability of the clusters
   c) the cohesion and separability of the clusters
   d) the non-convex shape of the clusters

2. A silhouette score indicates a high quality of clusters when the value is...
   a) ... close to 0.
   b) ... close to —1.
   c) ... larger than 1.
   d) ... close to 1.

3. Which of the following propositions is correct when a Gaussian mixture model is used?
    a) A data point has 1 as a membership value to one cluster and 0 for the other clusters.
    b) A data point has probability membership values to the different clusters.
    c) The provided clusters do not depend on the initialization.
    d) There is no need to define the number of clusters in advance.

LIST OF SOURCES

Images

Müller-Kett, 2020.
Müller-Kett, 2021.
Müller-Kett, 2023.
Microsoft Archive.