T6: Unsupervised learning

Fundamentos del Aprendizaje Automático

Curso 2025/2026

Structure

Outline

Problem statement

- **Supervised learning**: target to *guide* the learning process $\Rightarrow P(\omega|\mathbf{x})$

Problem statement

- **Supervised learning**: target to *guide* the learning process $\Rightarrow P(\omega|\mathbf{x})$
  - $\rightarrow$ Statistical distributions $P(\mathbf{x}|\omega)$, $P(\omega)$
  - $\rightarrow$ Data assortment $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$

Problem statement

- **Supervised learning**: target to *guide* the learning process $\Rightarrow P(\omega|\mathbf{x})$
  - $\rightarrow$ Statistical distributions $P(\mathbf{x}|\omega)$, $P(\omega)$
  - $\rightarrow$ Data assortment $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$

- **Unsupervised learning**: No specific target is provided
  - $\rightarrow$ Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$
  - $\rightarrow$ Related to Exploratory Data Analysis

Main tasks

1. **Clustering**:
   - Create groups (clusters) of data with similar characteristics
   - Identifying/characterizing patterns in the data

Main tasks

1. **Clustering**:
   - Create groups (clusters) of data with similar characteristics
   - Identifying/characterizing patterns in the data

2. **Dimensionality reduction**:
   - Derive compact representations of the initial feature space
   - Reduce complexity of the task
   - Visualization purposes

Main tasks

1. **Clustering**:
   - Create groups (clusters) of data with similar characteristics
   - Identifying/characterizing patterns in the data

2. **Dimensionality reduction**:
   - Derive compact representations of the initial feature space
   - Reduce complexity of the task
   - Visualization purposes

3. **Outlier/anomaly detection**:
   - Identify elements that *differ* for the rest in the assortment
   - Removenoise in assortments
   - Track unexpected behaviours

Outline

Definition

- Divide assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ into $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ clusters

Definition

- Divide assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ into $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ clusters

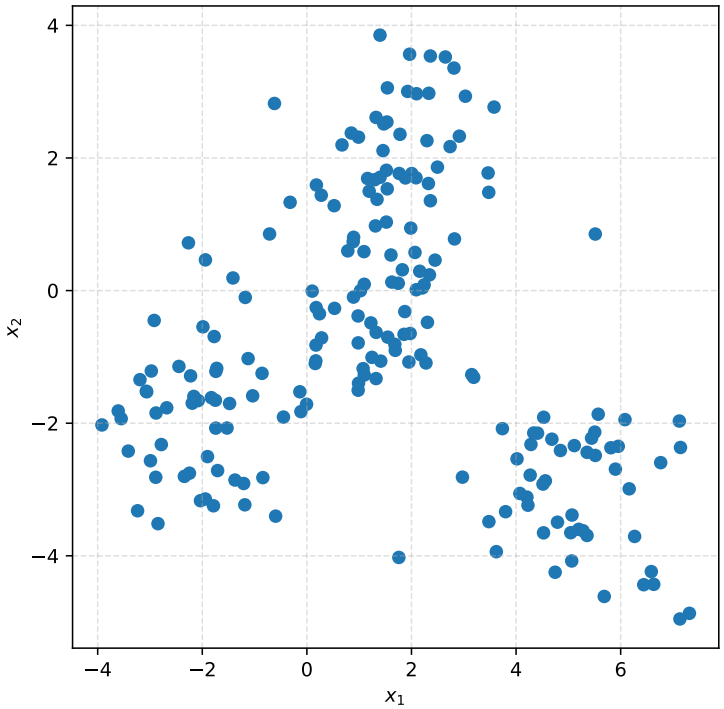
  $\Rightarrow \cup_{j=1}^{|\boldsymbol{\mathcal{C}}|} \mathcal{C}_j = \mathcal{D}$

Definition

- Divide assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ into $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ clusters
  $\Rightarrow \cup_{j=1}^{|\boldsymbol{\mathcal{C}}|} \mathcal{C}_j = \mathcal{D}$
- Division policy depends on ground principle $\Rightarrow$ Definition of *cluster*

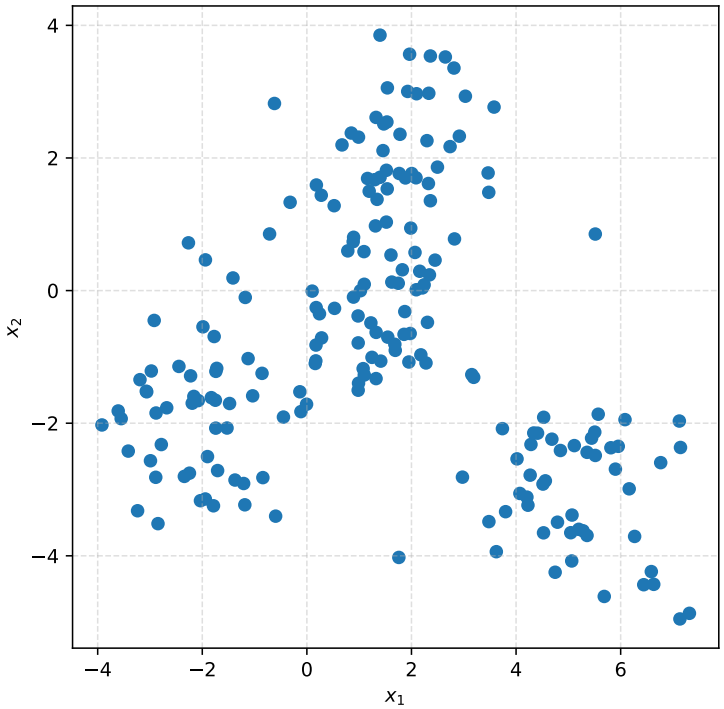
Definition

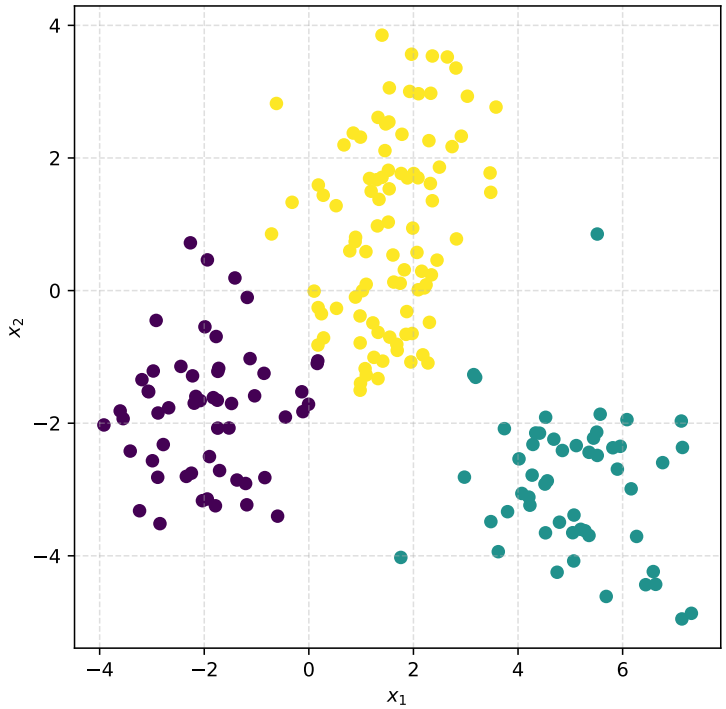
- Divide assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ into $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ clusters
  $\Rightarrow \cup_{j=1}^{|\boldsymbol{\mathcal{C}}|} \mathcal{C}_j = \mathcal{D}$
- Division policy depends on ground principle $\Rightarrow$ Definition of *cluster*



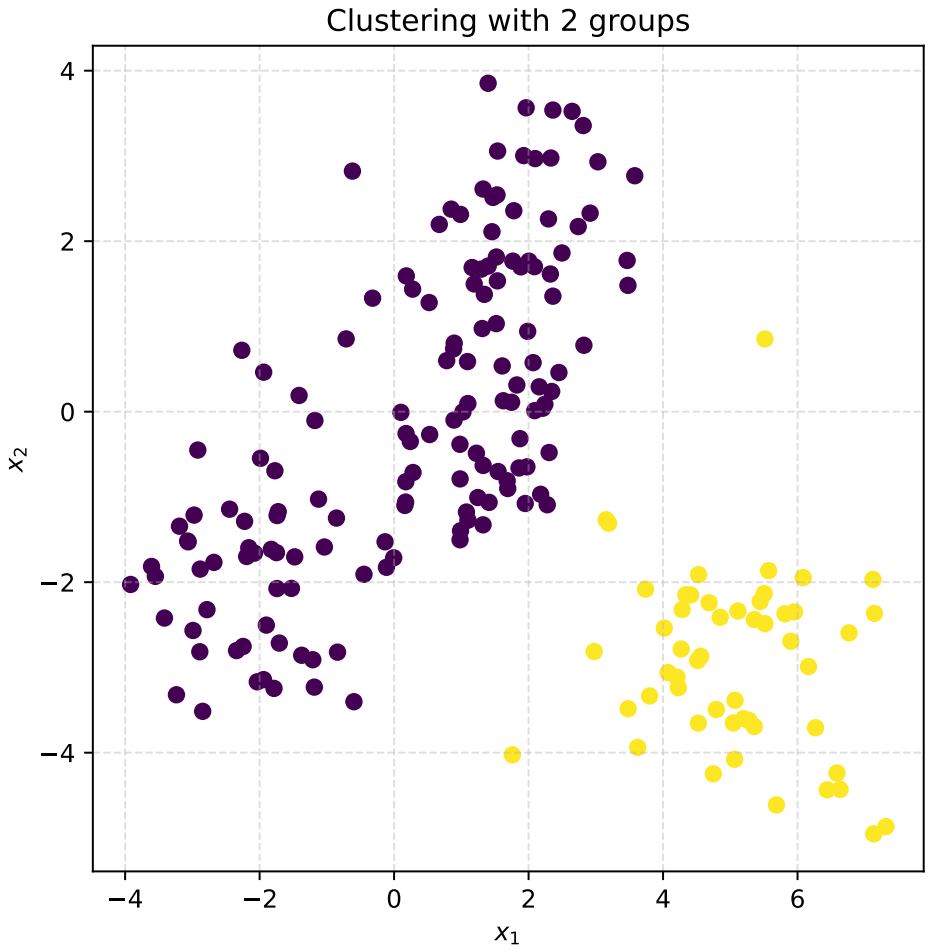
Definition

- Divide assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ into $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ clusters
  $\Rightarrow \cup_{j=1}^{|\boldsymbol{\mathcal{C}}|} \mathcal{C}_j = \mathcal{D}$
- Division policy depends on ground principle $\Rightarrow$ Definition of *cluster*





Example



Taxonomy

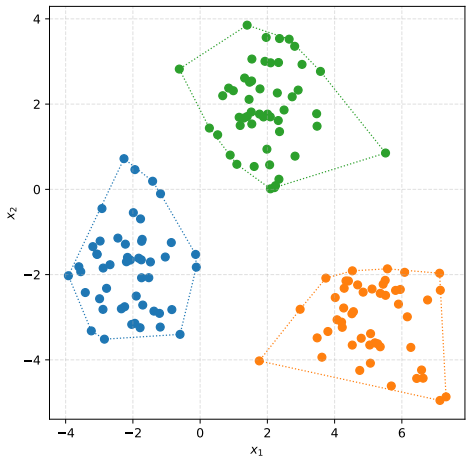
Taxonomy

1. Partitional clustering

2. Hierarchical clustering

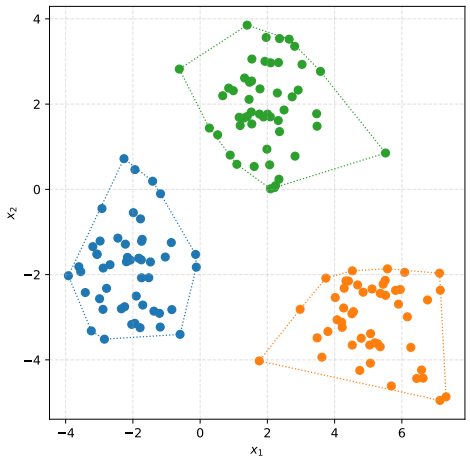
Taxonomy

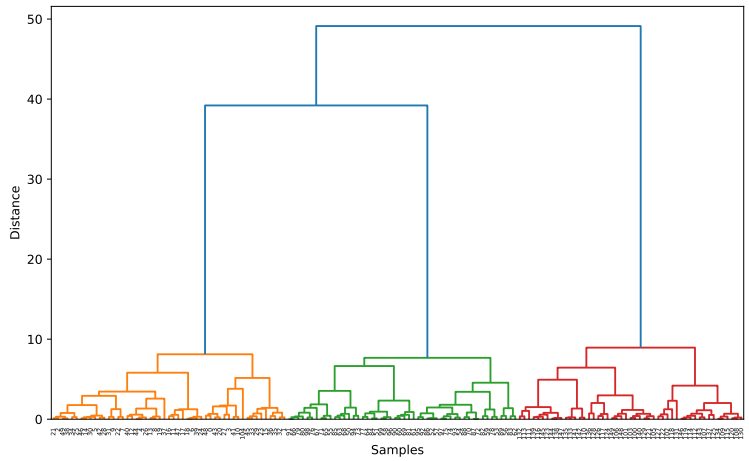
1. Partitional clustering
   - Data is organized in a plain structure without hierarchy
   - Generates $\mathcal{C}$ to recover the natural groupings inherent in $\mathcal{D}$
2. Hierarchical clustering



Taxonomy

1. Partitional clustering
   - Data is organized in a plain structure without hierarchy
   - Generates $\mathcal{C}$ to recover the natural groupings inherent in $\mathcal{D}$
2. Hierarchical clustering
   - $\mathcal{D}$ is partitioned into levels in a hierarchical format $\Rightarrow$ Dendrogram
     - $\rightarrow$ Tree-like diagram that shows how items are grouped
   - Two main approaches:
     - $\rightarrow$ **Top-down** or divisive
     - $\rightarrow$ **Bottom-up** or agglomerative





Partitional clustering

1.1 Hard clustering

1.2 Mixture resolving

1.3 Fuzzy clustering

Partitional clustering

1.1 Hard clustering

- Disjoint clusters: $\cap_{j=1}^{|\mathcal{C}|} \mathcal{C}_j = \emptyset$
- Common approaches: Graph-theoretic, density-based, centroid-based

1.2 Mixture resolving

1.3 Fuzzy clustering

Partitional clustering

1.1 Hard clustering

- Disjoint clusters: $\cap_{j=1}^{|\mathcal{C}|} \mathcal{C}_j = \emptyset$
- Common approaches: Graph-theoretic, density-based, centroid-based

1.2 Mixture resolving

- Assumes points come from mixtures of statistical distributions
- Goal is to estimate the parameters
- Example: Gaussian Mixture Model

1.3 Fuzzy clustering

Partitional clustering

1.1 Hard clustering

- Disjoint clusters: $\cap_{j=1}^{|\mathcal{C}|} \mathcal{C}_j = \emptyset$
- Common approaches: Graph-theoretic, density-based, centroid-based

1.2 Mixture resolving

- Assumes points come from mixtures of statistical distributions
- Goal is to estimate the parameters
- Example: Gaussian Mixture Model

1.3 Fuzzy clustering

- Clusters are defined as *fuzzy* sets
- Elements may belong to more than one cluster: $\cap_{j=1}^{|\mathcal{C}|} \mathcal{C}_j \neq \emptyset$

Definition

- Centroid/distance-based method $\Rightarrow$ Partitional and hard clustering

Definition

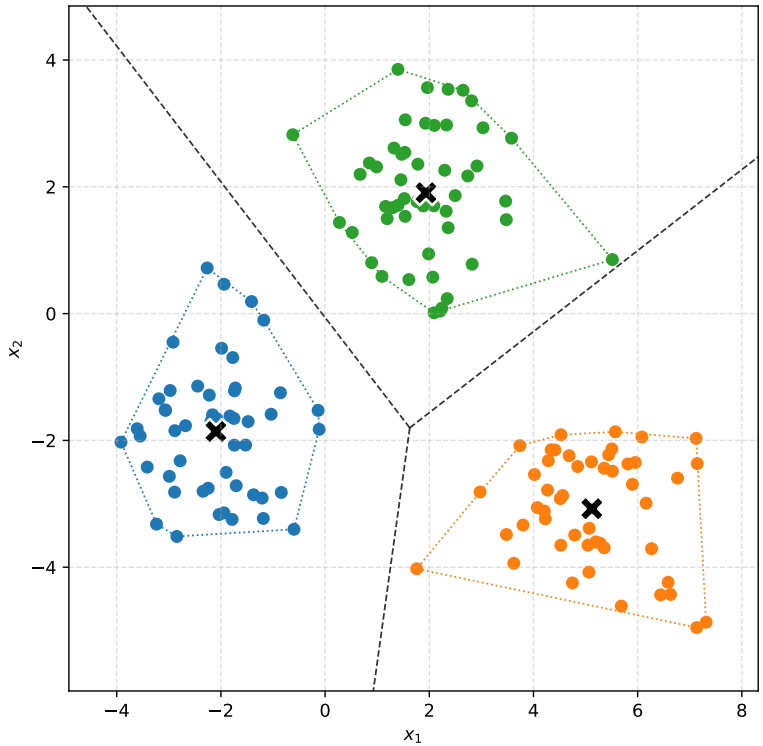
- Centroid/distance-based method $\Rightarrow$ Partitional and hard clustering
  - $\rightarrow$ Somehow k-Nearest Neighbor for unsupervised learning

Definition

- Centroid/distance-based method $\Rightarrow$ Partitional and hard clustering
  - $\rightarrow$ Somehow k-Nearest Neighbor for unsupervised learning
- Cluster: subset of elements whose **distance** to its **centroid** is minimal

Definition

- Centroid/distance-based method $\Rightarrow$ Partitional and hard clustering
  - $\rightarrow$ Somehow k-Nearest Neighbor for unsupervised learning
- Cluster: subset of elements whose **distance** to its **centroid** is minimal



Formulation

We assume the following conditions:

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$
- Distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$
- Distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$
- Total amount of $|\mathcal{C}|$ clusters $\Rightarrow \mathcal{C} = \{\mathcal{C}_j\}_{j=1}^{|\mathcal{C}|}$

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$
- Distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$
- Total amount of $|\mathcal{C}|$ clusters $\Rightarrow \mathcal{C} = \{\mathcal{C}_j\}_{j=1}^{|\mathcal{C}|}$
  - $\rightarrow$ Collection of centroids $\{\boldsymbol{\mu}_j\}_{j=1}^{|\mathcal{C}|}$ where $\boldsymbol{\mu}_j \in \mathbb{R}^d$

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$
- Distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$
- Total amount of $|\mathcal{C}|$ clusters $\Rightarrow \mathcal{C} = \{\mathcal{C}_j\}_{j=1}^{|\mathcal{C}|}$
  - $\rightarrow$ Collection of centroids $\{\boldsymbol{\mu}_j\}_{j=1}^{|\mathcal{C}|}$ where $\boldsymbol{\mu}_j \in \mathbb{R}^d$
  - $\rightarrow$ Centroid: typically, mean of its elements $\Rightarrow \boldsymbol{\mu}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_j} \mathbf{x}$

Formulation

We assume the following conditions:

- Data assortment $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i \in \mathbb{R}^d$
- Distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$
- Total amount of $|\mathcal{C}|$ clusters $\Rightarrow \mathcal{C} = \{\mathcal{C}_j\}_{j=1}^{|\mathcal{C}|}$
  - $\rightarrow$ Collection of centroids $\{\boldsymbol{\mu}_j\}_{j=1}^{|\mathcal{C}|}$ where $\boldsymbol{\mu}_j \in \mathbb{R}^d$
  - $\rightarrow$ Centroid: typically, mean of its elements $\Rightarrow \boldsymbol{\mu}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_j} \mathbf{x}$

Elements from $\mathcal{D}$ associated to a given cluster $\mathcal{C}_j$:

$$\mathcal{C}_j = \left\{ \mathbf{x}_i \in \mathcal{D} : \underset{1 \leq m \leq |\mathcal{C}|}{\arg\min}\, d\left(\mathbf{x}_i, \boldsymbol{\mu}_m\right) \right\}$$

Formulation

Premise: Given a number of clusters $|\mathcal{C}|$ and a distance $d(\cdot, \cdot)$, obtain the data agrupation that minimizes the accumulated distance to the centroids

Formulation

Premise: Given a number of clusters $|\mathcal{C}|$ and a distance $d(\cdot, \cdot)$, obtain the data agrupation that minimizes the accumulated distance to the centroids

$$\arg\min_{\mathcal{C}} \underbrace{\sum_{j=1}^{|\mathcal{C}|} \sum_{\mathbf{x} \in \mathcal{C}_j} d(\mathbf{x}, \boldsymbol{\mu}_j)}_{\textit{Within-Cluster Sum of Squares}}$$

Formulation

Premise: Given a number of clusters $|\mathcal{C}|$ and a distance $d(\cdot, \cdot)$, obtain the data agrupation that minimizes the accumulated distance to the centroids

$$\arg\min_{\mathcal{C}} \underbrace{\sum_{j=1}^{|\mathcal{C}|} \sum_{\mathbf{x} \in \mathcal{C}_j} d(\mathbf{x}, \boldsymbol{\mu}_j)}_{\textit{Within-Cluster Sum of Squares}}$$

The exact solution to this equation is NP-hard

- The iterative Lloyd's algorithm provides a suboptimal solution

Lloyd's algorithm

Iterative algorithm that alternates between two steps:

Lloyd's algorithm

Iterative algorithm that alternates between two steps:

1. **Assignment** step:

2. **Update** step:

Lloyd's algorithm

Iterative algorithm that alternates between two steps:

1. **Assignment** step:
   - Assign each datum $\mathbf{x}_i \in \mathcal{D}$ to the corresponding cluster $\mathcal{C}_j \in \boldsymbol{\mathcal{C}}$:

$$\mathcal{C}_j = \left\{ \mathbf{x}_i \in \mathcal{D} : \underset{1 \leq m \leq |\boldsymbol{\mathcal{C}}|}{\arg\min}\, d\left(\mathbf{x}_i, \boldsymbol{\mu}_m\right) \right\}$$

2. **Update** step:

Lloyd's algorithm

Iterative algorithm that alternates between two steps:

1. **Assignment** step:
   - Assign each datum $\mathbf{x}_i \in \mathcal{D}$ to the corresponding cluster $\mathcal{C}_j \in \boldsymbol{\mathcal{C}}$:

$$\mathcal{C}_j = \left\{ \mathbf{x}_i \in \mathcal{D} : \underset{1 \leq m \leq |\boldsymbol{\mathcal{C}}|}{\arg\min}\, d\left(\mathbf{x}_i, \boldsymbol{\mu}_m\right) \right\}$$

2. **Update** step:
   - Compute the centroid values $\boldsymbol{\mu}_j$ with $1 \leq j \leq |\boldsymbol{\mathcal{C}}|$:

$$\boldsymbol{\mu}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_j} \mathbf{x}$$

Lloyd's algorithm

Iterative algorithm that alternates between two steps:

1. **Assignment** step:
   - Assign each datum $\mathbf{x}_i \in \mathcal{D}$ to the corresponding cluster $\mathcal{C}_j \in \boldsymbol{\mathcal{C}}$:

$$\mathcal{C}_j = \left\{ \mathbf{x}_i \in \mathcal{D} : \underset{1 \leq m \leq |\boldsymbol{\mathcal{C}}|}{\arg\min}\, d\left(\mathbf{x}_i, \boldsymbol{\mu}_m\right) \right\}$$

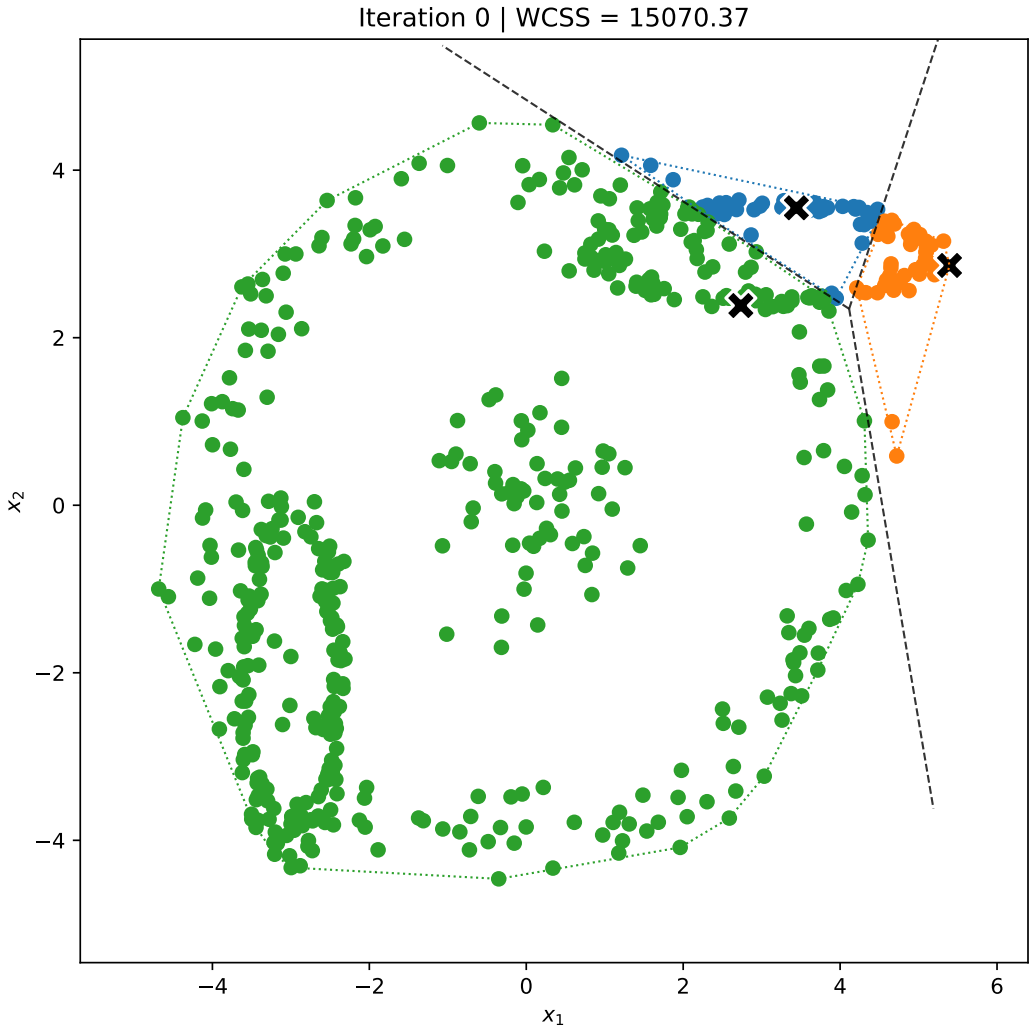
2. **Update** step:
   - Compute the centroid values $\boldsymbol{\mu}_j$ with $1 \leq j \leq |\boldsymbol{\mathcal{C}}|$:

$$\boldsymbol{\mu}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_j} \mathbf{x}$$

Process seeks to minimize *Within-Cluster Sum of Squares*:

- No changes in the cluster assigment between iterations

Lloyd's algorithm



Initialization

Initialization plays a key role in the convergence:

Initialization

Initialization plays a key role in the convergence:

- *Classic* k-means: selects $|\mathcal{C}|$ random points from $\mathcal{D}$ as centroids

Initialization

Initialization plays a key role in the convergence:

- *Classic* k-means: selects $|\mathcal{C}|$ random points from $\mathcal{D}$ as centroids
- k-means++: avoids issues related to random initialization

Initialization

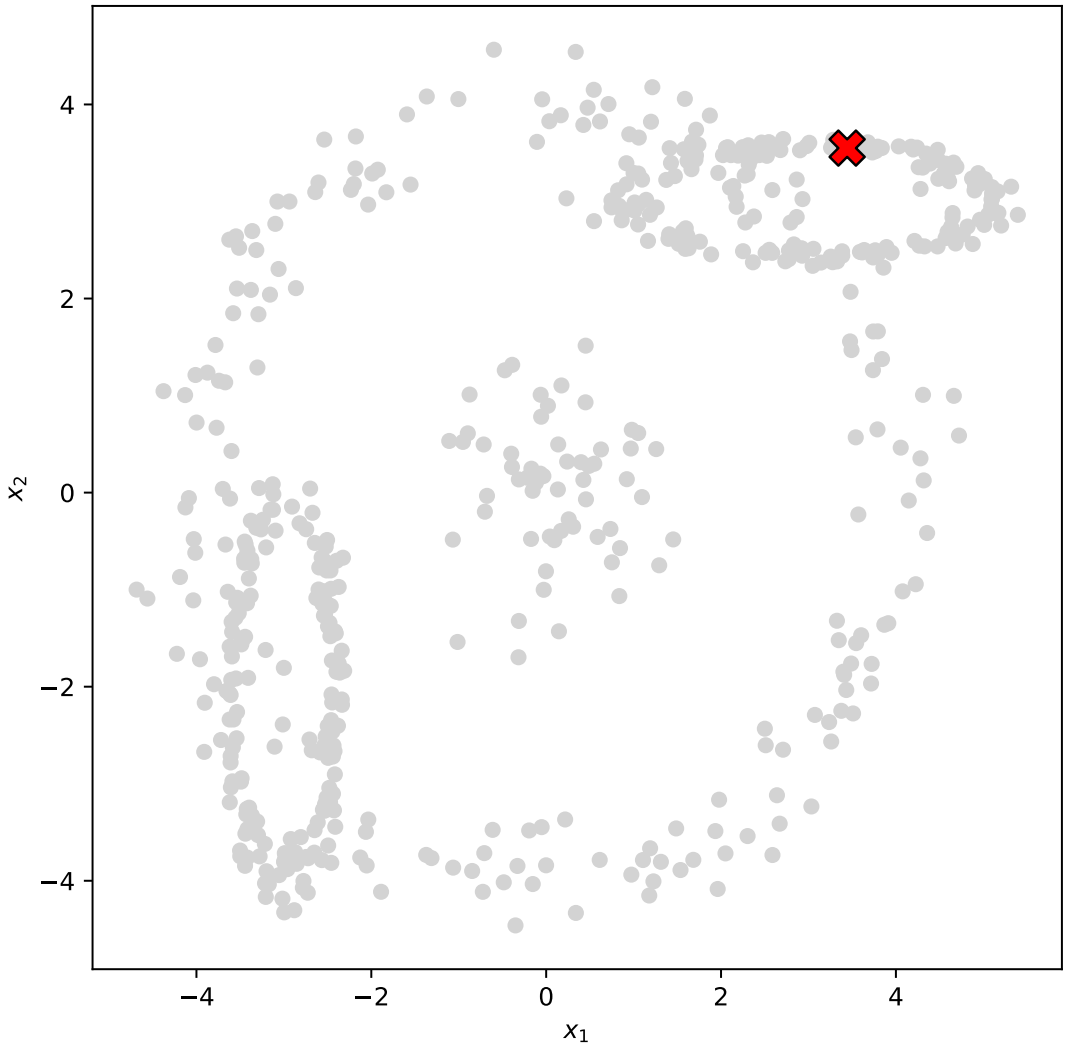
Initialization plays a key role in the convergence:

- *Classic* k-means: selects $|\mathcal{C}|$ random points from $\mathcal{D}$ as centroids
- k-means++: avoids issues related to random initialization

k-means++ initialization method

Input: Number of clusters $|\mathcal{C}|$
Result: Set of centroids $\{\boldsymbol{\mu}_j\}_{j=1}^{|\mathcal{C}|}$
$K = 1$, $\boldsymbol{\mu}_1 \in_R \mathcal{D}$;
while $K < |\mathcal{C}|$ **do**
  $\mathcal{P} = \left\{\min_{k=1}^{K} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2\right\} \forall \mathbf{x}_i \in \mathcal{D}$;
  $\boldsymbol{\mu}_{K+1} = \mathcal{D}_{\text{Categorical}(\mathcal{P}/\sum \mathcal{P})}$;
  $K = K + 1$;
end

Example of k-means++



Additional remarks

✓ May address both statistical and structural representations
- Only requires a distance metric

Additional remarks

✓ May address both statistical and structural representations
- Only requires a distance metric

✗ Mean operator is sensitive to outliers

Additional remarks

✓ May address both statistical and structural representations
- Only requires a distance metric

✗ Mean operator is sensitive to outliers

✓ Alternative approach: k-Medoids
- *In-set mean* $\Rightarrow$ Closest element in cluster $\mathcal{C}_j$ to mean $\boldsymbol{\mu}_j$ $(1 \leq j \leq |\boldsymbol{\mathcal{C}}|)$

$$\boldsymbol{\mu}_j^s = \arg\min_{\mathbf{x} \in \mathcal{C}_j} d\left(\mathbf{x}, \boldsymbol{\mu}_j\right)$$

Contextualization

Key question $\Rightarrow$ how many clusters do we have in our data?

Contextualization

Key question $\Rightarrow$ how many clusters do we have in our data?

- ✗ It is not possible to know the *true* number of groups (unsupervised)

Contextualization

Key question $\Rightarrow$ how many clusters do we have in our data?

- ✗ It is not possible to know the *true* number of groups (unsupervised)
- ✓ There exist strategies to estimate this parameter

Contextualization

Key question $\Rightarrow$ how many clusters do we have in our data?

- ✗ It is not possible to know the *true* number of groups (unsupervised)
- ✓ There exist strategies to estimate this parameter

Two representative strategies to perform this estimation:

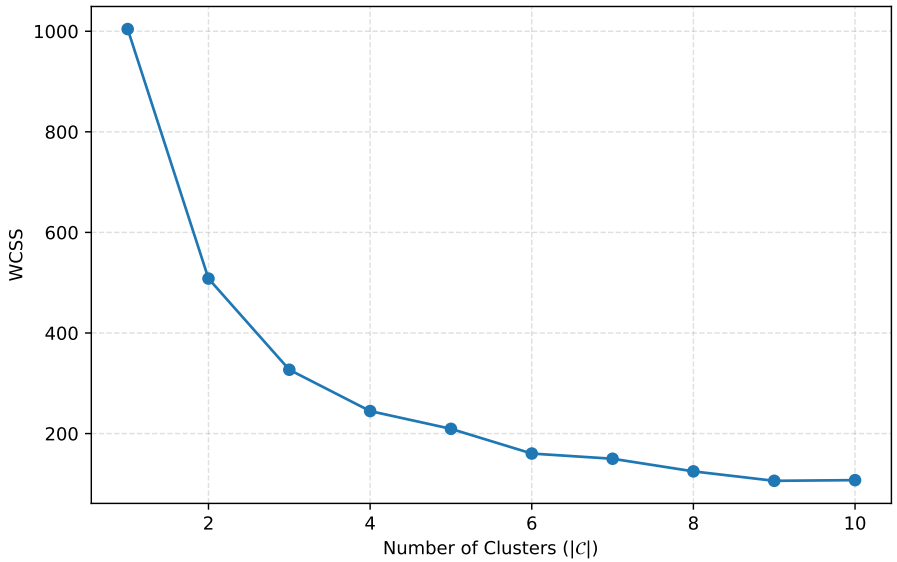
1. Elbow method
2. Silhoutte method

Elbow method

- In general, higher $|\mathcal{C}|$ values imply lower WCSS scores

Elbow method

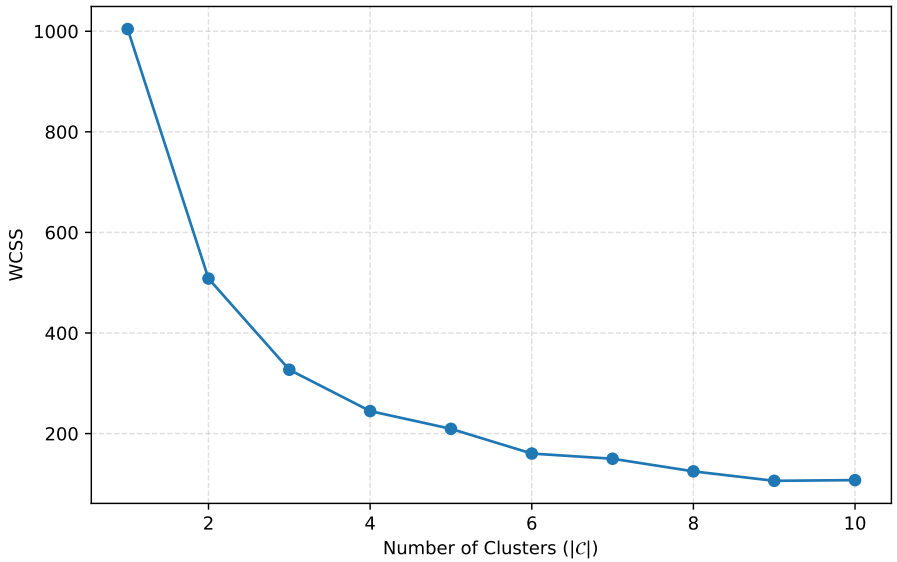
- In general, higher $|\mathcal{C}|$ values imply lower WCSS scores
  - $\rightarrow$ Graphically, this relation resembles a negative exponential curve



Elbow method

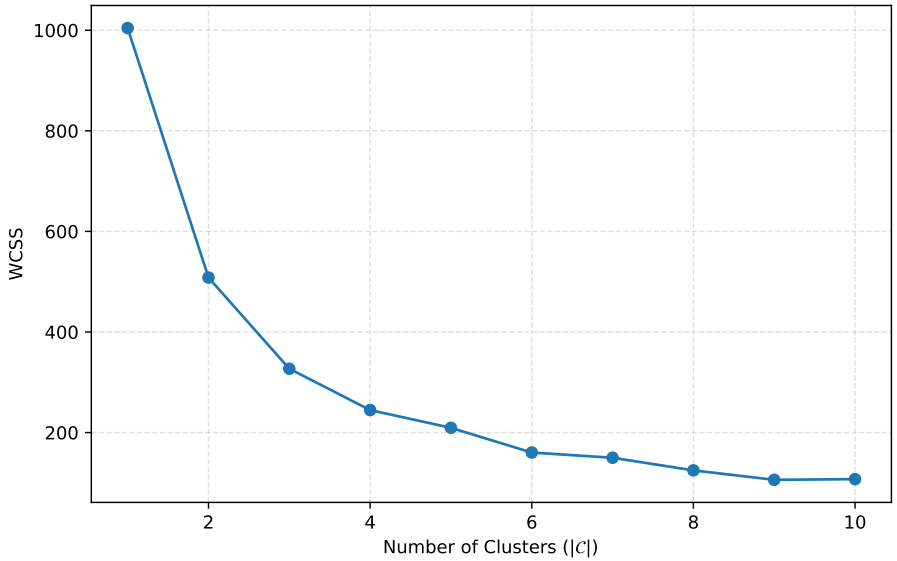
- In general, higher $|\mathcal{C}|$ values imply lower WCSS scores
  - $\rightarrow$ Graphically, this relation resembles a negative exponential curve

- After a certain point: more clusters $\Leftrightarrow$ small WCSS improvements
  - $\rightarrow$ *Turning* point $\Rightarrow$ Elbow of the curve

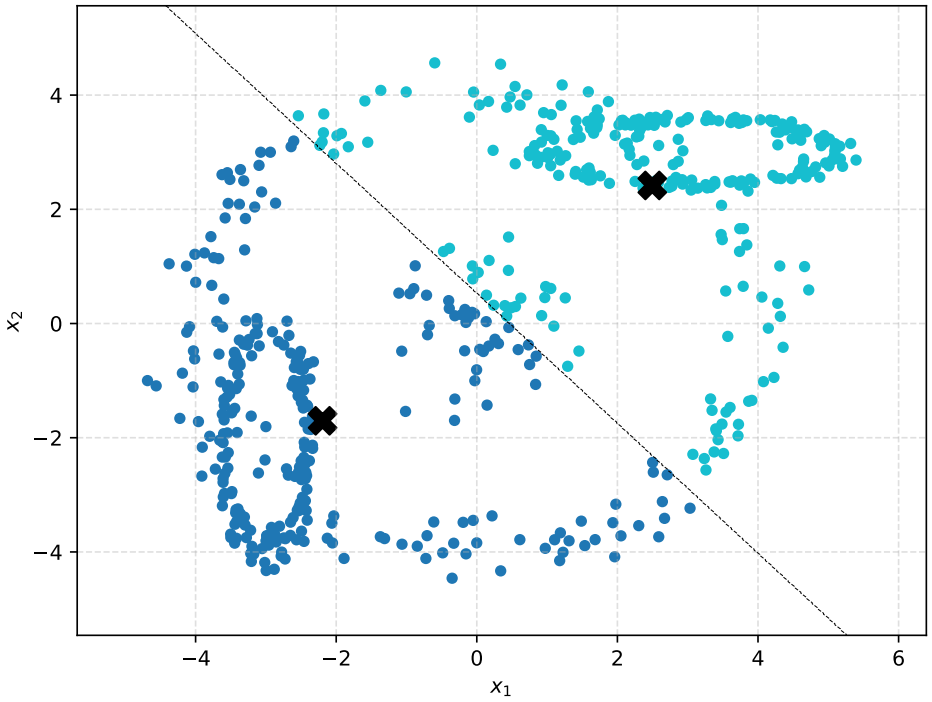


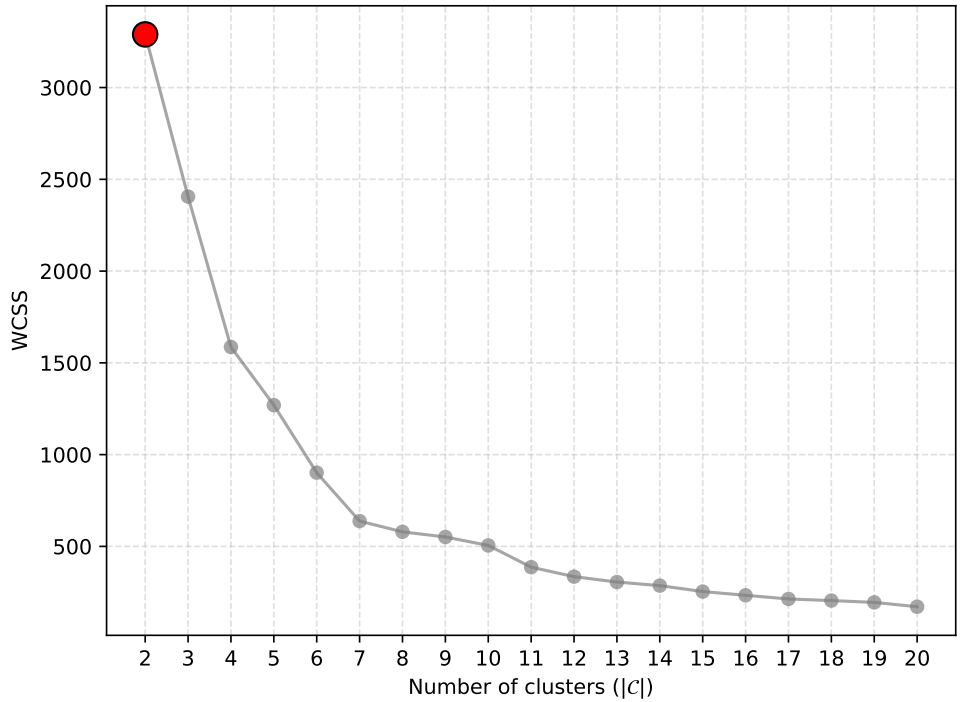
Elbow method

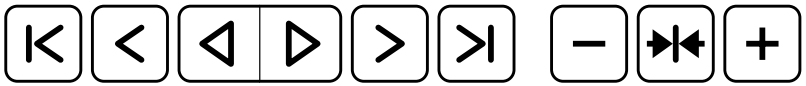
- In general, higher $|\mathcal{C}|$ values imply lower WCSS scores
  - → Graphically, this relation resembles a negative exponential curve
- After a certain point: more clusters ⇔ small WCSS improvements
  - → *Turning* point ⇒ Elbow of the curve
- The elbow is assumed to be the optimal number of clusters $|\mathcal{C}|$



Elbow method







Silhouette method

Measures how well each sample $\mathbf{x}_i$ fits into its assigned cluster $\mathcal{C}_j$ compared to other clusters

Silhouette method

Measures how well each sample $\mathbf{x}_i$ fits into its assigned cluster $\mathcal{C}_j$ compared to other clusters

Considers two aspects for assessing the *quality* of the clustering:

1. Compactness: how tight a cluster is
2. Separation: how far apart clusters are

Silhouette method

Consider a point $\mathbf{x}_m \in \mathcal{D}$ currently assigned to cluster $\mathcal{C}_j$:

Silhouette method

Consider a point $\mathbf{x}_m \in \mathcal{D}$ currently assigned to cluster $\mathcal{C}_j$:

- Cohesion indicator $a(\mathbf{x}_m)$:

$$a(\mathbf{x}_m) = \frac{1}{|\mathcal{C}_j| - 1} \sum_{\mathbf{x}_n \in \mathcal{C}_j, \mathbf{x}_m \neq \mathbf{x}_n} d(\mathbf{x}_m, \mathbf{x}_n)$$

Silhouette method

Consider a point $\mathbf{x}_m \in \mathcal{D}$ currently assigned to cluster $\mathcal{C}_j$:

- Cohesion indicator $a(\mathbf{x}_m)$:

$$a\left(\mathbf{x}_m\right) = \frac{1}{|\mathcal{C}_j| - 1} \sum_{\mathbf{x}_n \in \mathcal{C}_j, \mathbf{x}_m \neq \mathbf{x}_n} d\left(\mathbf{x}_m, \mathbf{x}_n\right)$$

- Separation indicator $b(\mathbf{x}_m)$:

$$b\left(\mathbf{x}_m\right) = \min_{c \in \{1,\ldots,|\boldsymbol{\mathcal{C}}|\}, c \neq j} \frac{1}{|\mathcal{C}_c|} \sum_{\mathbf{x}_n \in \mathcal{C}_c} d\left(\mathbf{x}_m, \mathbf{x}_n\right)$$

Silhouette method

Consider a point $\mathbf{x}_m \in \mathcal{D}$ currently assigned to cluster $\mathcal{C}_j$:

- Cohesion indicator $a(\mathbf{x}_m)$:

$$a\left(\mathbf{x}_m\right) = \frac{1}{|\mathcal{C}_j| - 1} \sum_{\mathbf{x}_n \in \mathcal{C}_j, \mathbf{x}_m \neq \mathbf{x}_n} d\left(\mathbf{x}_m, \mathbf{x}_n\right)$$

- Separation indicator $b(\mathbf{x}_m)$:

$$b\left(\mathbf{x}_m\right) = \min_{c \in \{1,\ldots,|\boldsymbol{\mathcal{C}}|\}, c \neq j} \frac{1}{|\mathcal{C}_c|} \sum_{\mathbf{x}_n \in \mathcal{C}_c} d\left(\mathbf{x}_m, \mathbf{x}_n\right)$$

- Silhouette score $s(\mathbf{x}_m)$:

$$s(\mathbf{x}_m) = \frac{b(\mathbf{x}_m) - a(\mathbf{x}_n)}{\max\left\{a(\mathbf{x}_m), b(\mathbf{x}_m)\right\}} \in [-1, +1]$$

Silhouette method

The Silhouette indicator exhibits three situations:

$$s(\mathbf{x}_m) \begin{cases} > 0 \text{ element adequately clustered} \\ = 0 \text{ element on decision boundary} \\ < 0 \text{ element probably in incorrect cluster} \end{cases}$$

Silhouette method

The Silhouette indicator exhibits three situations:

$$s(\mathbf{x}_m) \begin{cases} > 0 \text{ element adequately clustered} \\ = 0 \text{ element on decision boundary} \\ < 0 \text{ element probably in incorrect cluster} \end{cases}$$

Since the indicator is individual, a global Silhouette may be derived as:

$$\hat{s} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_m \in \mathcal{D}} s(\mathbf{x}_m)$$

Silhouette method

The Silhouette indicator exhibits three situations:

$$s(\mathbf{x}_m) \begin{cases} > 0 \text{ element adequately clustered} \\ = 0 \text{ element on decision boundary} \\ < 0 \text{ element probably in incorrect cluster} \end{cases}$$

Since the indicator is individual, a global Silhouette may be derived as:

$$\hat{s} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_m \in \mathcal{D}} s(\mathbf{x}_m)$$

$\rightarrow$ Can be used as alternative metric for the Elbow method

Silhouette method