University of Southern Denmark
IMADA

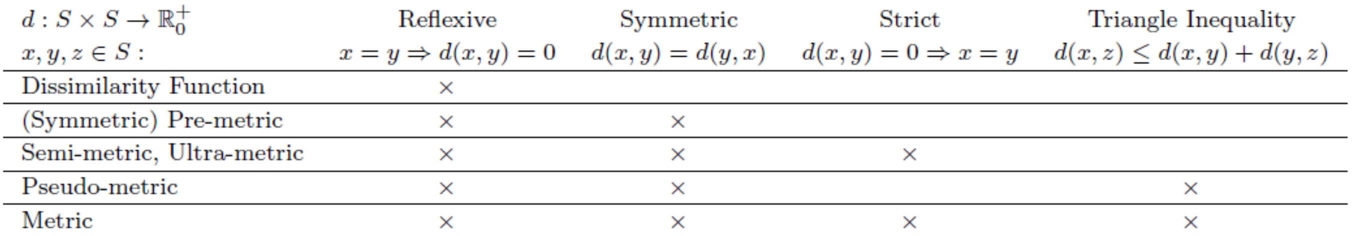
DM566: Data Mining and Machine Learning

Spring term 2022

Exercise 4: Distance Measures, Clustering, Silhouette

Exercise 4-1 Distance functions (1 point)

Distance functions can be classified into the following categories:



$d : S \times S \to \mathbb{R}_0^+$	Reflexive	Symmetric	Strict	Triangle Inequality
$x, y, z \in S$:	$x = y \Rightarrow d(x,y) = 0$	$d(x,y) = d(y,x)$	$d(x,y) = 0 \Rightarrow x = y$	$d(x,z) \leq d(x,y) + d(y,z)$
Dissimilarity Function	×			
(Symmetric) Pre-metric	×	×		
Semi-metric, Ultra-metric	×	×	×	
Pseudo-metric	×	×		×
Metric	×	×	×	×

Decide for each of the following functions $d(\mathbb{R}^n, \mathbb{R}^n)$ whether they are a distance function, and if so, which type.

1. $d(x,y) = \sum_{i=1}^{n}(x_i - y_i)$
   Suggested solution:
   As this function can take negative numbers, it does not satisfy $d : S \times S \to \mathbb{R}_0^+$.

2. $d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$
   Suggested solution:
   As this function is squared, it satisfies $d : S \times S \to \mathbb{R}_0^+$.
   It is reflexive, symmetric and strict.

   The triangle inequality is not satisfied.
   Counter example: $o = (0,0)$, $p = (1,0)$, $q = (2,0)$:

$$
\begin{gathered}
d(o,q) = 4 \\
d(o,p) + d(p,q) = 1 + 1 = 2 \\
4 > 2
\end{gathered}
$$

3. $d(x,y) = \sqrt{\sum_{i=1}^{n-1}(x_i - y_i)^2}$
   Suggested solution:
   As this function is squared, it satisfies $d : S \times S \to \mathbb{R}_0^+$.
   It is reflexive, symmetric and satisfies the triangle inequality.
   It is not strict, as it only sums up to $n - 1$, so $d(x,y) = 0$ is possible, while $x \neq y$.

4. $d(x, y) = \sum_{i=1}^{n}\{1 \text{ iff } x_i = y_i, \quad 0 \text{ iff } x_i \neq y_i\}$

   Suggested solution:

   It satisfies $d : S \times S \rightarrow \mathbb{R}_0^+$.

   It is not reflexive – the other properties are therefore irrelevant to us.

5. $d(x, y) = \sum_{i=1}^{n}\{0 \text{ iff } x_i = y_i, \quad 1 \text{ iff } x_i \neq y_i\}$

   Suggested solution:

   It satisfies $d : S \times S \rightarrow \mathbb{R}_0^+$.

   It is reflexive, symmetric and strict.

   It satisfies triangle inequality.

   Proof of triangle inequality by case distinction on the individual positions:

   (i) $x_i = y_i \wedge y_i = z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 0 + 0 &\geq 0 \end{aligned}$$

   (ii) $x_i = y_i \wedge x_i \neq z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(x_i, z_i) &\geq d(x_i, z_i) \\ 0 + 1 &\geq 1 \end{aligned}$$

   (iii) $x_i = z_i \wedge x_i \neq y_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, y_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 1 + 1 &\geq 0 \end{aligned}$$

   (iv) $x_i \neq y_i \wedge y_i = z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, y_i) + d(y_i, y_i) &\geq d(x_i, y_i) \\ 1 + 0 &\geq 1 \end{aligned}$$

   (v) $x_i \neq y_i \wedge y_i \neq z_i \wedge x_i \neq z_i$:

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ 1 + 1 &\geq 1 \end{aligned}$$

Exercise 4-2 k-means, 1-dimensional example (1 point)

Given the following 1-dimensional datapoints: $\{2, 3, 4, 10, 11, 12, 20, 25, 30\}$.
For $k = \{2, 3, 4\}$ and sets of initial means $\{\mu_1 = 2, \mu_2 = 6\}$, $\{\mu_1 = 2, \mu_2 = 4, \mu_3 = 6\}$, $\{\mu_1 = 2, \mu_2 = 4, \mu_3 = 6, \mu_4 = 10\}$, compute the new clusters after each iteration of k-means (Lloyd/Forgy) until convergence.

Suggested solution:

For $k = 2$:

- Starting with the initial means $\{\mu_1 = 2, \mu_2 = 6\}$, we assign each point to the closest mean, which yields the following clusters:
  $C_1 = \{2, 3, 4\}$
  $C_2 = \{10, 11, 12, 20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 3$
  $\mu_2 = 18$

- For the second iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3, 4, 10\}$
  $C_2 = \{11, 12, 20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 4.75$
  $\mu_2 = 19.6$

- For the third iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3, 4, 10, 11, 12\}$
  $C_2 = \{20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 7$
  $\mu_2 = 25$

  Thereafter, the clusters do not change.

For $k = 3$:

- Starting with the initial means $\{\mu_1 = 2, \mu_2 = 4$ and $\mu_3 = 6\}$, we assign each point to the closest mean, which yields the following clusters:
  $C_1 = \{2, 3\}$
  $C_2 = \{4\}$
  $C_3 = \{10, 11, 12, 20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 2.5$
  $\mu_2 = 4$
  $\mu_3 = 18$

- For the second iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3\}$
  $C_2 = \{4, 10, 11\}$
  $C_3 = \{12, 20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 2.5$
  $\mu_2 = 8.33$
  $\mu_3 = 21.75$

- For the third iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3, 4\}$
  $C_2 = \{10, 11, 12\}$
  $C_3 = \{20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 3$
  $\mu_2 = 11$
  $\mu_3 = 25$

  Thereafter, the clusters do not change.

For $k = 4$:

- Starting with the initial means $\{\mu_1 = 2, \mu_2 = 4, \mu_3 = 6\}$ and $\mu_4 = 10$, we assign each point to the closest mean, which yields the following clusters:
  $C_1 = \{2, 3\}$
  $C_2 = \{4\}$
  $C_3 = \{\}$
  $C_4 = \{10, 11, 12, 20, 25, 30\}$

  Note that since C_3 is empty, we can simply ignore this cluster and continue the next iteration with $k - 1$ clusters.
  The new means are as follows:
  $\mu_1 = 2.5$
  $\mu_2 = 4$
  $\mu_3 =?$
  $\mu_4 = 18$

  The previous C_4 and μ_4 will from this point on be called C_3 and μ_3.

- For the second iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3\}$
  $C_2 = \{4, 10, 11\}$
  $C_3 = \{12, 20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 2.5$
  $\mu_2 = 8.33$
  $\mu_3 = 21.75$

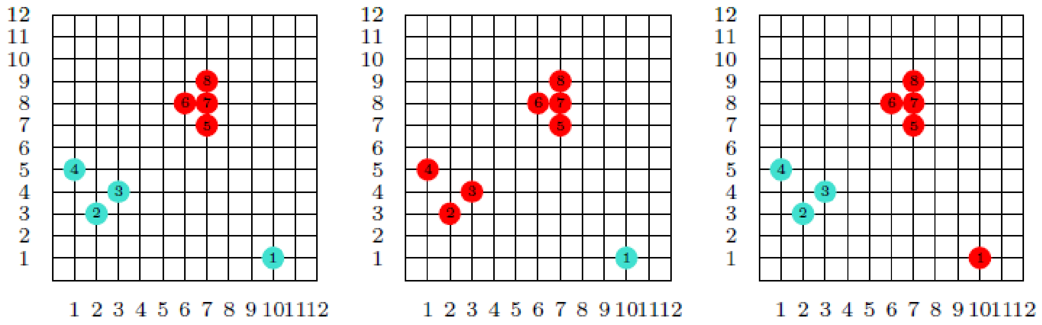
- For the third iteration, the assignment to the closest mean yields the following clusters:
  $C_1 = \{2, 3, 4\}$
  $C_2 = \{10, 11, 12\}$
  $C_3 = \{20, 25, 30\}$

  The new means are as follows:
  $\mu_1 = 3$
  $\mu_2 = 11$
  $\mu_3 = 25$

  Thereafter, the clusters do not change.

Exercise 4-3 Silhouette Coefficient (1 point)

We derived three different clustering solutions for the toy data set in the lecture:



Compute the simplified silhouette coefficient for each solution. Compare the result with the ranking by the k-means objective function (TD^2), that we determined in the lecture.

Suggested solution:

Recall how the simplified silhouette coefficient is calculated.

Let $a(o)$ be the distance between o and its “own” cluster representative.

Let $b(o)$ be the distance between o and the closest “foreign” cluster representative.

The silhouette of o is given by

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

where $s(o) \approx -1, 0, 1$: bad, indifferent, good assignment of o.

Solution 1: $\mu_{blue} = (4, 3.25)$, $\mu_{red} = (6.75, 8)$

$TD^2 = 61\frac{1}{2}$

$$s(p_1) = \frac{7.718 - 6.408}{\max(6.408, 7.718)} \approx 0.17$$

$$s(p_2) = \frac{6.897 - 2.016}{\max(2.016, 6.897)} \approx 0.708$$

$$s(p_3) = \frac{5.483 - 1.25}{\max(1.25, 5.483)} \approx 0.772$$

$$s(p_4) = \frac{6.486 - 3.473}{\max(3.473, 6.486)} \approx 0.464$$

$$s(p_5) = \frac{4.802 - 1.031}{\max(1.031, 4.802)} \approx 0.785$$

$$s(p_6) = \frac{5.154 - 0.75}{\max(0.75, 5.154)} \approx 0.854$$

$$s(p_7) = \frac{5.618 - 0.25}{\max(0.25, 5.618)} \approx 0.956$$

$$s(p_8) = \frac{6.486 - 1.031}{\max(1.031, 6.486)} \approx 0.841$$

Silhouette Coefficient for Clustering 1: 0.694

Solution 2: $\mu_{blue} = (10, 1)$, $\mu_{red} = (4.7, 6.3)$
$TD^2 = 72.68$

$$\begin{aligned}
s(p_1) &= \frac{7.475 - 0.0}{\max(0.0, 7.475)} \approx 1.0 \\
s(p_2) &= \frac{8.246 - 4.262}{\max(4.262, 8.246)} \approx 0.483 \\
s(p_3) &= \frac{7.616 - 2.857}{\max(2.857, 7.616)} \approx 0.625 \\
s(p_4) &= \frac{9.849 - 3.931}{\max(3.931, 9.849)} \approx 0.601 \\
s(p_5) &= \frac{6.708 - 2.395}{\max(2.395, 6.708)} \approx 0.643 \\
s(p_6) &= \frac{8.062 - 2.143}{\max(2.143, 8.062)} \approx 0.734 \\
s(p_7) &= \frac{7.616 - 2.857}{\max(2.857, 7.616)} \approx 0.625 \\
s(p_8) &= \frac{8.544 - 3.548}{\max(3.548, 8.544)} \approx 0.585
\end{aligned}$$

Silhouette Coefficient for Clustering 2: 0.662

Solution 3: $\mu_{blue} = (2, 4)$, $\mu_{red} = (7.4, 6.6)$
$TD^2 = 54\frac{2}{5}$

$$\begin{aligned}
s(p_1) &= \frac{8.544 - 6.174}{\max(6.174, 8.544)} \approx 0.277 \\
s(p_2) &= \frac{6.49 - 1.0}{\max(1.0, 6.49)} \approx 0.846 \\
s(p_3) &= \frac{5.111 - 1.0}{\max(1.0, 5.111)} \approx 0.804 \\
s(p_4) &= \frac{6.597 - 1.414}{\max(1.414, 6.597)} \approx 0.786 \\
s(p_5) &= \frac{5.831 - 0.566}{\max(0.566, 5.831)} \approx 0.903 \\
s(p_6) &= \frac{5.657 - 1.98}{\max(1.98, 5.657)} \approx 0.65 \\
s(p_7) &= \frac{6.403 - 1.456}{\max(1.456, 6.403)} \approx 0.773 \\
s(p_8) &= \frac{7.071 - 2.433}{\max(2.433, 7.071)} \approx 0.656
\end{aligned}$$

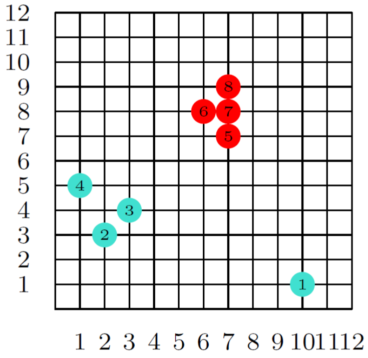
Silhouette Coefficient for Clustering 3: 0.712
We thus get the same ranking with simplified silhouette as for TD^2.

Exercise 4-4 Silhouette Coefficient (1 point)

Apply the Lloyd-Forgy variant of k-means clustering on the same data set as Exercise 4-3 but this time with three clusters with initial centers $\mu_1 = (9, 10), \mu_2 = (2, 5), \mu_3 = (12, 4)$. What is the silhouette coefficient of this clustering? Compare to the clusterings and silhouette coefficients in Exercise 3-4. Explain which cluster count you would choose and why.

Suggested solution:

We have the following dataset from the previous exercise.



- Starting with the initial centers $\mu_1 = (9, 10), \mu_2 = (2, 5), \mu_3 = (12, 4)$ of the 3 clusters, we assign each point to the closest center, which yields the following clusters:
  $C_1 = \{5, 6, 7, 8\}$
  $C_2 = \{2, 3, 4\}$
  $C_3 = \{1\}$

- We then recompute the centers:
  $\mu_1 = (6.75, 8)$
  $\mu_2 = (2, 4)$
  $\mu_3 = (10, 1)$

- We assign each point to the closest center again, and notice that there is no change in the assignment.

  The result of the algorithm is then the clustering written above.

We calculate the silhouette coefficient of this clustering as follows:

$$s(p_1) = \frac{7.718 - 0.0}{\max(0.0, 7.718)} \approx 1.0$$
$$s(p_2) = \frac{6.897 - 1.0}{\max(1.0, 6.897)} \approx 0.855$$
$$s(p_3) = \frac{5.483 - 1.0}{\max(1.0, 5.483)} \approx 0.818$$
$$s(p_4) = \frac{6.486 - 1.414}{\max(1.414, 6.486)} \approx 0.782$$
$$s(p_5) = \frac{5.831 - 1.031}{\max(1.031, 5.831)} \approx 0.823$$

$$s(p_6) = \frac{5.657 - 0.75}{\max(0.75, 5.657)} \approx 0.867$$
$$s(p_7) = \frac{6.403 - 0.25}{\max(0.25, 6.403)} \approx 0.961$$
$$s(p_8) = \frac{7.071 - 1.031}{\max(1.031, 7.071)} \approx 0.854$$

Silhouette Coefficient for Clustering: 0.87.
This Silhouette Coefficient is better than all the Silhouette Coefficients from exercise 3-4.