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:

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

So if a distance measure satisfies $d : S \times S \to \mathbb{R}_0^+$ and $\forall x, y, z \in S$, it is reflexive, symmetric, and strict, and it also satisfies the triangle inequality, then it is a metric.

As you can see, a pre-metric does not necessarily need to be *strictly* reflexive. Make sure you understand the difference between reflexivity and strictness.

Note: these terms as well as “distance function” are used inconsistently in the literature. In mathematics, “distance function” is commonly used synonymously with “metric”. In a database and data mining context, strictness is often not relevant at all, and a “distance function” usually refers to a pseudo-metric, pre-metric, or even just to some dissimilarity function. Do not rely on Wikipedia, it uses multiple definitions within itself!

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. $d(x,y) = \sum_{i=1}^{n}(x_i - y_i)$
2. $d(x,y) = \sum_{i=1}^{n}(x_i - y_1)^2$
3. $d(x,y) = \sqrt{\sum_{i=1}^{n-1}(x_i - y_i)^2}$
4. $d(x,y) = \sum_{i=1}^{n}\{1 \text{ iff } x_i = y_1, \quad 0 \text{ iff } x_i \neq y_1\}$
5. $d(x,y) = \sum_{i=1}^{n}\{0 \text{ iff } x_i = y_1, \quad 1 \text{ iff } x_i \neq y_1\}$

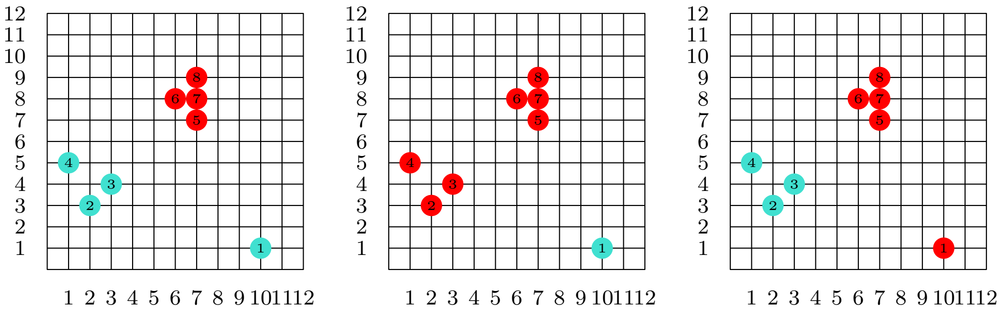
Exercise 4-2 k-means, 1-dimensional example (1 point)

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.

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.

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.