Exercise 4 : Distance Measures, Clustering, Silhouette

Exercise 4-1 : Distance functions

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.

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

(a) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)$

(b) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

(c) $d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$

(d) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i = y_i \\ 0 & \text{iff} \quad x_i \neq y_i \end{cases}$

(e) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$

Exercise 4-2 : Distances on a database

Given a database similar to this one :

r	x	y
1	0	1
2	1	1
3	0	1

r	x	y
4	1	1
5	2	2
6	3	3

Which properties does the following distance function have ?

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Explain which records are considered equivalent by this distance function, and discuss whether it is sensible in a database and data mining context to have pseudo-metric distance functions.

Hint : What could be the nature of attribute r in a database context ?

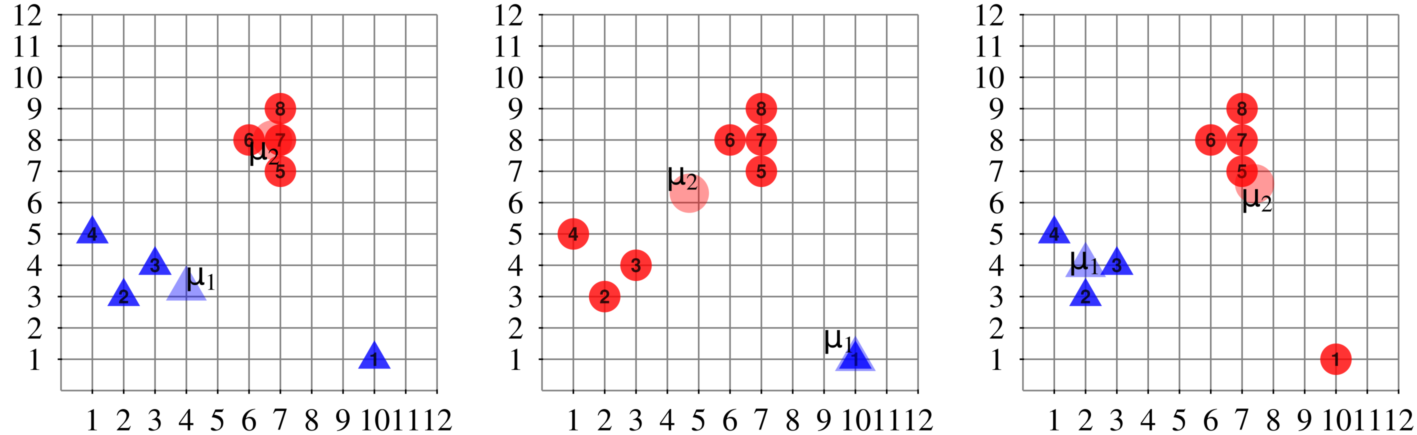
Exercise 4-3 : k-means 1-dimensional Example

Given are the following 1-dimensional points : $\{2, 3, 4, 10, 11, 12, 20, 25, 30\}$. We set $k = 3$ and choose as initial means : $\mu_1 = 2$, $\mu_2 = 4$, and $\mu_3 = 6$.

Compute the new clusters after each iteration of k-means (Lloyd/Forgy) until convergence.

Exercise 4-4 : Silhouette Coefficient

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 4-5 : Tools

K-Means clustering. Calculate Silhouette and TD^2 score and perform Silhouette analysis appropriately to determine best k for number of clusters using YellowBrick (a machine learning visualization library).

(a) Load python packages : `datasets, metrics` from `sklearn`, and `KMeans` from `sklearn.cluster`.

(b) Load wine dataset and assign data as x and target as y.

(c) Define and fit the kmeans model and check different number of clusters.

(d) Calculate Silhoutte and TD^2 Score and print them for different number of clusters.

(e) Load `SilhouetteVisualizer` from `yellowbrick.cluster`, and `matplotlib.pyplot`.

(f) Create `Silhouette` plot for `K-Means` cluster with different Methods of initialization (`"k-means++"`, `"random"`). Then check various number of clusters : 2, 3, 4, 5.

(g) What is the most appropriate K for number of clusters.