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

Suggested solution :
d is not even positive definite : it can become negative.

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

Suggested solution :
d is reflexive, symmetric, strict, but the triangle inequality is not satisfied.
Counter example : $o = (0, 0)$, $p = (1, 0)$, $q = (2, 0)$:

$$d(o, q) = 4 \geq 1 + 1 = d(o, p) + d(p, q)$$

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

Suggested solution :
d is reflexive, symmetric, satisfies the triangle inequality, but is not strict.

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

Suggested solution :
d is not reflexive – the other properties are therefore irrelevant to us.

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

Suggested solution :

d defines the so-called Hamming distance, a metric which plays an important role in information theory. On binary representations, it corresponds to the number of ones after an XOR operation of two binary vectors.

Reflexivity, symmetry and strictness should be obvious.

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

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

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

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

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

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

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

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

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

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

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

Which implies :

$$\begin{aligned} d(x,y)+d(y,z) &= \sum_{i}^{n} d(x_i,y_i) + \sum_{i}^{n} d(y_i,z_i) \\ &= \sum_{i}^{n} \left(d(x_i,y_i)+d(y_i,z_i)\right) \\ &\geq \sum_{i}^{n} d(x_i,z_i) \\ &= d(x,z) \end{aligned}$$

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 ?

Suggested solution :

If we ignore r, which is just the record number (key) in the database, this is Euclidean distance on $\mathbb{R}^2$ – known to be a metric on $\mathbb{R}^2$. All metric properties *except* strictness are the same, but obviously records 1 and 3 are not the same, while having a distance of 0.

Records are equivalent if and only if they have the same "coordinates", i.e., they are "duplicate records", and differ only by the record number r !

Many "metric" distance functions will turn out to be only pseudo-metrics when used in databases. We have to take care of this anyway.

Note that many distance functions work on derived features, where different database objects could be mapped to the same point in the feature space.

Example : two different images that have the same proportional amount of colors (perhaps given some reduction of the color space) would be mapped to the same point in the feature space defined by the distribution of colors (i.e., they have identical color histograms). In the feature space, any distance measure would indicate equality (i.e., we could infer identity) of the two images (and in fact, their representations in this feature space are identical) although the original objects (images) that are represented in that feature space are not the same.

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.

Suggested solution :

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 :

$C_1 = \{2, 3\}$,

$C_3 = \{10, 11, 12, 20, 25, 30\}$,

$C_2 = \{4\}$,

where we assigned 3 to C_1 instead of C_2 (arbitrary resolvement of a tie).

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 :

$\mu_1 = 3$,

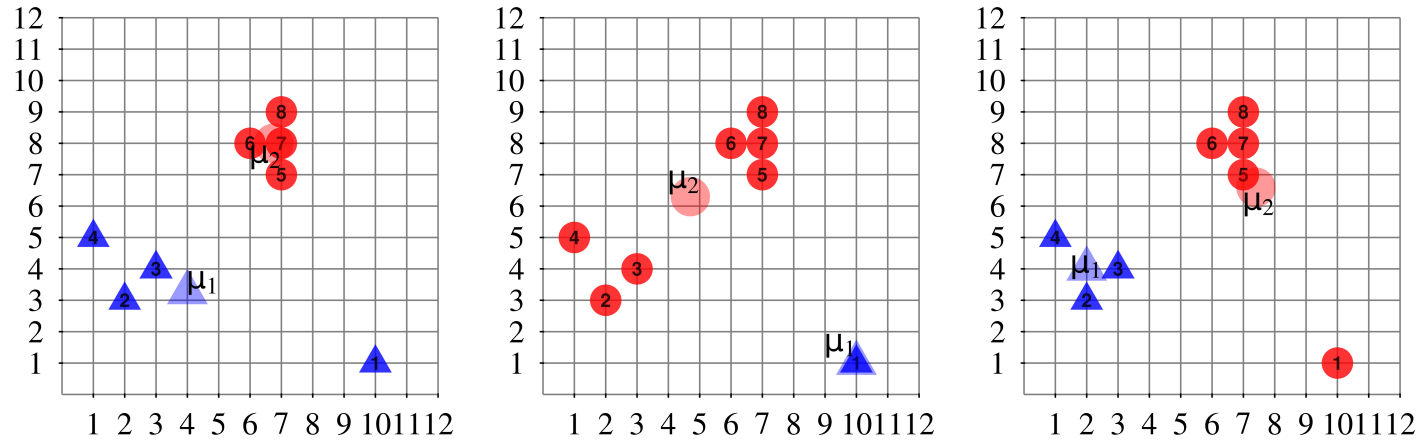
$\mu_2 = 11$,

$\mu_3 = 25$.

Thereafter, the clusters do not change.

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.

Suggested solution :

Solution 1 : $\mu_1 = (4, 3.25)$, $\mu_2 = (6.75, 8)$

$dist(\mu_1, p_1) = \sqrt{|4.0-10|^2 + |3.25-1|^2} = 6.408$ $dist(\mu_2, p_1) = \sqrt{|6.75-10|^2 + |8.0-1|^2} = 7.718$

$dist(\mu_1, p_2) = \sqrt{|4.0-2|^2 + |3.25-3|^2} = 2.016$ $dist(\mu_2, p_2) = \sqrt{|6.75-2|^2 + |8.0-3|^2} = 6.897$

$dist(\mu_1, p_3) = \sqrt{|4.0-3|^2 + |3.25-4|^2} = 1.25$ $dist(\mu_2, p_3) = \sqrt{|6.75-3|^2 + |8.0-4|^2} = 5.483$

$dist(\mu_1, p_4) = \sqrt{|4.0-1|^2 + |3.25-5|^2} = 3.473$ $dist(\mu_2, p_4) = \sqrt{|6.75-1|^2 + |8.0-5|^2} = 6.486$

$dist(\mu_2, p_5) = \sqrt{|6.75-7|^2 + |8.0-7|^2} = 1.031$ $dist(\mu_1, p_5) = \sqrt{|4.0-7|^2 + |3.25-7|^2} = 4.802$

$dist(\mu_2, p_6) = \sqrt{|6.75-6|^2 + |8.0-8|^2} = 0.75$ $dist(\mu_1, p_6) = \sqrt{|4.0-6|^2 + |3.25-8|^2} = 5.154$

$dist(\mu_2, p_7) = \sqrt{|6.75-7|^2 + |8.0-8|^2} = 0.25$ $dist(\mu_1, p_7) = \sqrt{|4.0-7|^2 + |3.25-8|^2} = 5.618$

$dist(\mu_2, p_8) = \sqrt{|6.75-7|^2 + |8.0-9|^2} = 1.031$ $dist(\mu_1, p_8) = \sqrt{|4.0-7|^2 + |3.25-9|^2} = 6.486$

Calculating Silhouettes

$$s(p_1) = \frac{7.718 - 6.408}{\max(6.408, 7.718)} \approx 0.17 \qquad s(p_5) = \frac{4.802 - 1.031}{\max(1.031, 4.802)} \approx 0.785$$

$$s(p_2) = \frac{6.897 - 2.016}{\max(2.016, 6.897)} \approx 0.708 \qquad s(p_6) = \frac{5.154 - 0.75}{\max(0.75, 5.154)} \approx 0.854$$

$$s(p_3) = \frac{5.483 - 1.25}{\max(1.25, 5.483)} \approx 0.772 \qquad s(p_7) = \frac{5.618 - 0.25}{\max(0.25, 5.618)} \approx 0.956$$

$$s(p_4) = \frac{6.486 - 3.473}{\max(3.473, 6.486)} \approx 0.464 \qquad s(p_8) = \frac{6.486 - 1.031}{\max(1.031, 6.486)} \approx 0.841$$

Silhouette Coefficient for Clustering 1 : 0.694

Solution 2 : $\mu_1 = (10, 1)$, $\mu_2 = (4.7, 6.3)$

$dist(\mu_1, p_1) = \sqrt{|10.0 - 10|^2 + |1.0 - 1|^2} = 0.0$ $dist(\mu_2, p_1) = \sqrt{|4.714 - 10|^2 + |6.286 - 1|^2} = 7.475$

$dist(\mu_2, p_2) = \sqrt{|4.714 - 2|^2 + |6.286 - 3|^2} = 4.262$ $dist(\mu_1, p_2) = \sqrt{|10.0 - 2|^2 + |1.0 - 3|^2} = 8.246$
$dist(\mu_2, p_3) = \sqrt{|4.714 - 3|^2 + |6.286 - 4|^2} = 2.857$ $dist(\mu_1, p_3) = \sqrt{|10.0 - 3|^2 + |1.0 - 4|^2} = 7.616$
$dist(\mu_2, p_4) = \sqrt{|4.714 - 1|^2 + |6.286 - 5|^2} = 3.931$ $dist(\mu_1, p_4) = \sqrt{|10.0 - 1|^2 + |1.0 - 5|^2} = 9.849$
$dist(\mu_2, p_5) = \sqrt{|4.714 - 7|^2 + |6.286 - 7|^2} = 2.395$ $dist(\mu_1, p_5) = \sqrt{|10.0 - 7|^2 + |1.0 - 7|^2} = 6.708$
$dist(\mu_2, p_6) = \sqrt{|4.714 - 6|^2 + |6.286 - 8|^2} = 2.143$ $dist(\mu_1, p_6) = \sqrt{|10.0 - 6|^2 + |1.0 - 8|^2} = 8.062$
$dist(\mu_2, p_7) = \sqrt{|4.714 - 7|^2 + |6.286 - 8|^2} = 2.857$ $dist(\mu_1, p_7) = \sqrt{|10.0 - 7|^2 + |1.0 - 8|^2} = 7.616$
$dist(\mu_2, p_8) = \sqrt{|4.714 - 7|^2 + |6.286 - 9|^2} = 3.548$ $dist(\mu_1, p_8) = \sqrt{|10.0 - 7|^2 + |1.0 - 9|^2} = 8.544$

Calculating Silhouettes

$$s(p_1) = \frac{7.475 - 0.0}{\max(0.0, 7.475)} \approx 1.0$$

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

Silhouette Coefficient for Clustering 2 : 0.662

Solution 3 : $\mu_1 = (2, 4)$, $\mu_2 = (7.4, 6.6)$

$dist(\mu_1, p_2) = \sqrt{|2.0-2|^2 + |4.0-3|^2} = 1.0$

$dist(\mu_1, p_3) = \sqrt{|2.0-3|^2 + |4.0-4|^2} = 1.0$

$dist(\mu_1, p_4) = \sqrt{|2.0-1|^2 + |4.0-5|^2} = 1.414$

$dist(\mu_2, p_1) = \sqrt{|7.4-10|^2 + |6.6-1|^2} = 6.174$

$dist(\mu_2, p_5) = \sqrt{|7.4-7|^2 + |6.6-7|^2} = 0.566$

$dist(\mu_2, p_6) = \sqrt{|7.4-6|^2 + |6.6-8|^2} = 1.98$

$dist(\mu_2, p_7) = \sqrt{|7.4-7|^2 + |6.6-8|^2} = 1.456$

$dist(\mu_2, p_9) = \sqrt{|7.4-7|^2 + |6.6-9|^2} = 2.433$

$dist(\mu_2, p_2) = \sqrt{|7.4-2|^2 + |6.6-3|^2} = 6.49$

$dist(\mu_2, p_3) = \sqrt{|7.4-3|^2 + |6.6-4|^2} = 5.111$

$dist(\mu_2, p_4) = \sqrt{|7.4-1|^2 + |6.6-5|^2} = 6.597$

$dist(\mu_1, p_1) = \sqrt{|2.0-10|^2 + |4.0-1|^2} = 8.544$

$dist(\mu_1, p_5) = \sqrt{|2.0-7|^2 + |4.0-7|^2} = 5.831$

$dist(\mu_1, p_6) = \sqrt{|2.0-6|^2 + |4.0-8|^2} = 5.657$

$dist(\mu_1, p_7) = \sqrt{|2.0-7|^2 + |4.0-8|^2} = 6.403$

$dist(\mu_1, p_9) = \sqrt{|2.0-7|^2 + |4.0-9|^2} = 7.071$

Calculating Silhouettes

$$s(p_1) = \frac{8.544 - 6.174}{\max(6.174, 8.544)} \approx 0.277$$

$$s(p_2) = \frac{6.49 - 1.0}{\max(1.0, 6.49)} \approx 0.846$$

$$s(p_3) = \frac{5.111 - 1.0}{\max(1.0, 5.111)} \approx 0.804$$

$$s(p_4) = \frac{6.597 - 1.414}{\max(1.414, 6.597)} \approx 0.786$$

$$s(p_5) = \frac{5.831 - 0.566}{\max(0.566, 5.831)} \approx 0.903$$

$$s(p_6) = \frac{5.657 - 1.98}{\max(1.98, 5.657)} \approx 0.65$$

$$s(p_7) = \frac{6.403 - 1.456}{\max(1.456, 6.403)} \approx 0.773$$

$$s(p_9) = \frac{7.071 - 2.433}{\max(2.433, 7.071)} \approx 0.656$$

Silhouette Coefficient for Clustering 3 : 0.712

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`.

Suggested solution :

```
from sklearn import datasets
from sklearn import metrics
from sklearn.cluster import KMeans
```

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

Suggested solution :

```
# Load Wine dataset
wine = datasets.load_wine()
X = wine.data
y = wine.target
```

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

Suggested solution :

```
# Instantiate the KMeans models
km = KMeans(n_clusters=3, random_state=42)
# Fit the KMeans model
km.fit_predict(X)
```

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

Suggested solution :

```
# Calculate Silhoutte Score
score = metrics.silhouette_score(X, km.labels_, metric='euclidean')
# Print the Silhouetter score
print('Silhouetter Score: %.3f' % score)
# Calculate and Print TD2 the score
print('TD2 Score: %.3f' % km.inertia_)
```

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

Suggested solution :

```
from yellowbrick.cluster import SilhouetteVisualizer
import matplotlib.pyplot as plt
```

(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.

Suggested solution :

```
#initialization methods
for init in ['k-means++', 'random']:
    fig, ax = plt.subplots(2, 2, figsize=(15,8))
    fig.suptitle(f'with initializatin method: {init}')
    for i in [2, 3, 4, 5]:
        #Create KMeans instance for different number of clusters
        km = KMeans(n_clusters=i, init=init, n_init=10,
        max_iter=100, random_state=42)
        q, mod = divmod(i, 2)
        #Create SilhouetteVisualizer instance with KMeans instance,
        #fit the visualizer
        visualizer = SilhouetteVisualizer(km, colors='yellowbrick',
        ax=ax[q-1][mod])
        visualizer.fit(X)
        ax[q-1][mod].set_title(f'TD2 score: {visualizer.inertia_}')
```

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