

COMPSCI 762 2022 S1 Week 12 Solutions

Luke Chang

June 7, 2022

Question 1 – Clustering

Figure 1: A data set has 8 instances, from A1 to A8. The Euclidean distances between every two instances are given as the following:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Use the data in Figure 1 to answer the questions below:

1.1

Suppose that the initial seeds (centers of each cluster) are **A1**, **A4** and **A7**. Run the k-means algorithm for 1 epoch only. At the end of this epoch show the new clusters (i.e. the examples belonging to each cluster)

Answer:

C1 is centred at **A1**, **C2** is centred at **A4**, and **C3** is centred at **A7**.

For each point:

- **A1** → **C1**

Table 1: For A2

Item	Centroid	Dist
A2	A1	$\sqrt{45}$
A2	A4	$\sqrt{49}$
A2	A7	$\sqrt{5}$

Table 2: For A3

Item	Centroid	Dist
A3	A1	$\sqrt{63}$
A3	A4	$\sqrt{11}$
A3	A7	$\sqrt{47}$

- **A2 \rightarrow C3**
- **A3 \rightarrow C2**
- **A4 \rightarrow C2**

Table 3: For A5

Item	Centroid	Dist
A5	A1	$\sqrt{41}$
A5	A4	$\sqrt{2}$
A5	A7	$\sqrt{21}$

- **A5 \rightarrow C2**

Table 4: For A6

Item	Centroid	Dist
A6	A1	$\sqrt{28}$
A6	A4	$\sqrt{7}$
A6	A7	$\sqrt{13}$

- **A6 \rightarrow C2**
- **A7 \rightarrow C3**

Table 5: For A8

Item	Centroid	Dist
A8	A1	$\sqrt{6}$
A8	A4	$\sqrt{5}$
A8	A7	$\sqrt{53}$

- **A8 \rightarrow C2**

The new clusters: $C1 = \{A1\}$, $C2 = \{A3, A4, A5, A6, A8\}$, $C3 = \{A2, A7\}$

1.2

Use single-linkage (MIN) agglomerative clustering to group the data. Show the dendrogram.

Answer:

Table 6: Step 1

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Table 7: Step 1 – Clusters

Level	# Clusters	Clusters
0	8	$\{A1\}, \{A2\}, \{A3\}, \{A4\}, \{A5\}, \{A6\}, \{A7\}, \{A8\}$
1	7	$\{A1\}, \{A2\}, \{A3\}, \{A4, A5\}, \{A6\}, \{A7\}, \{A8\}$

Table 8: Step 2

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Table 9: Step 2 – Clusters

Level	# Clusters	Clusters
0	8	$\{A1\}, \{A2\}, \{A3\}, \{A4\}, \{A5\}, \{A6\}, \{A7\}, \{A8\}$
1	7	$\{A1\}, \{A2\}, \{A3\}, \{A4, A5\}, \{A6\}, \{A7\}, \{A8\}$
2	4	$\{A1\}, \{A2, A7\}, \{A3\}, \{A4, A5, A6, A8\}$

Table 10: Step 3

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Table 11: Step 3 – Clusters

Level	# Clusters	Clusters
0	8	$\{A1\}, \{A2\}, \{A3\}, \{A4\}, \{A5\}, \{A6\}, \{A7\}, \{A8\}$
1	7	$\{A1\}, \{A2\}, \{A3\}, \{A4, A5\}, \{A6\}, \{A7\}, \{A8\}$
2	4	$\{A1\}, \{A2, A7\}, \{A3\}, \{A4, A5, A6, A8\}$
3	3	$\{A1, A4, A5, A6, A8\}, \{A2, A7\}, \{A3\}$

Table 12: Step 4

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Table 13: Step 4 – Clusters. Note that A4 and A6 are already in one cluster.

Level	# Clusters	Clusters
0	8	$\{A1\}, \{A2\}, \{A3\}, \{A4\}, \{A5\}, \{A6\}, \{A7\}, \{A8\}$
1	7	$\{A1\}, \{A2\}, \{A3\}, \{A4, A5\}, \{A6\}, \{A7\}, \{A8\}$
2	4	$\{A1\}, \{A2, A7\}, \{A3\}, \{A4, A5, A6, A8\}$
3	3	$\{A1, A4, A5, A6, A8\}, \{A2, A7\}, \{A3\}$
4	1	$\{A1, A2, A3, A4, A5, A6, A7, A8\}$

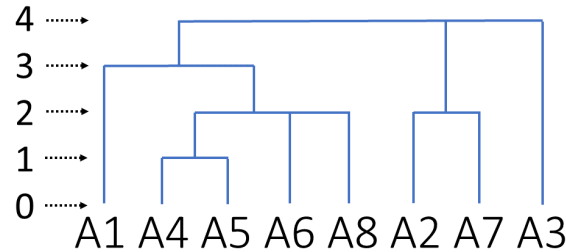


Figure 2: Step 5 – Dendrogram: Make sure you choose the correct order to start with. Use the sequence from the second last level;

Question 2 – Ooutlier/Anomaly Detection

2.1

What are the three types of anomaly? Give an example for each type.

Answer:

- **Global outlier (Point anomaly):** deviates significantly from the rest of the data set. The simplest type of outliers.
- **Contextual outlier (Conditional outlier):** deviates significantly with respect to a specific context of the object.
 - **Contextual attributes:** define the object's context.
 - **Behavioral attributes:** define the object's characteristics, and are used to evaluate whether the object is an outlier in the context.

For example:

A temperature sensor measures 4°C in May. It is a perfectly normal reading in Wellington, but it might be an outlier if the location is New York. The location and the date are **contextual attributes**, and the temperature is a **behavioral attribute**.

- **Collective outliers:** the objects as a whole deviate significantly from the entire data set.

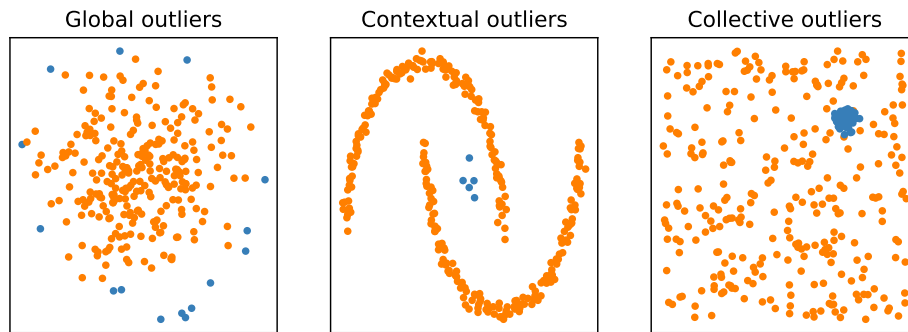


Figure 3: Types of outliers

Examples:

- Fig. 3 left: Global outliers
- Fig. 3 middle: Contextual outliers – Given the dataset has two clusters, each one has a moon shape.
- Fig. 3 right: Blue points are collective outliers because the density of those points is much higher than the rest.

2.2

You are given the following list of 2D data points:

$$[1; 1]; [1; 2]; [2; 2]; [2; 1]; [3; 3]; [2; 5]; [2; 3]$$

If you had to select one point to be anomalous, how to use Manhattan distance to determine the outlier. Explain the anomaly detection technique.

Answer:

There are multiple ways to solve this problem. Let's use **distance-based outlier detection**.

	1,1	1,2	2,2	2,1	3,3	2,5	2,3
1,1	0	1	2	1	4	5	3
1,2	1	0	1	2	3	4	2
2,2	2	1	0	1	2	3	1
2,1	1	2	1	0	3	4	2
3,3	4	3	2	3	0	3	1
2,5	5	4	3	4	3	0	2
2,3	3	2	1	2	1	2	0

Table 14: Manhattan Distance Matrix

$$\frac{|\{o' | \text{dist}(o, o') \leq r\}|}{|D|} \leq \pi$$

where $|D|$ is the *cardinality* of the set D .

We need to define the hyperparameters: r and π .

The number of data points, $|D|$, is 7.

Let $r = 2$,

	# of objects within r	Divide by $ D $
1,1	3	0.43
1,2	4	0.57
2,2	5	0.71
2,1	4	0.57
3,3	2	0.29
2,5	1	0.14
2,3	5	0.71

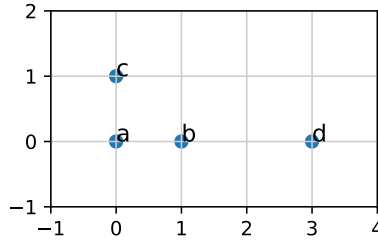
Table 15: The # of neighbours within r

If we select only one point as an outlier, we can set π to any value between 0.14 and 0.29, e.g. 0.15. Therefore, we identify $[2; 5]$ is an outlier.

2.3

Consider a set of points $(0,0)$, $(1,0)$, $(0,1)$, $(3,0)$. Calculate the *Local Outlier Factor* (LOF) score for the points using Manhattan distance and k is 2.

Answer:



	a (0,0)	b (1,0)	c (0,1)	d (3,0)
a (0,0)	0	1	1	3
b (1,0)	1	0	2	2
c (0,1)	1	2	0	4
d (3,0)	3	2	4	0

Table 16: Manhattan distance matrix

	$\text{dist}_2(o)$	$N_2(o)$
a (0,0)	1	2
b (1,0)	2	3
c (0,1)	2	2
d (3,0)	3	2

Table 17: Distance between data point o and its k -th nearest neighbour ($k = 2$)

We denote the set of k -nearest neighbours as $N_k(o)$:

$$N_k(o) = \{o' | o' \in D, \text{dist}(o, o') \leq \text{dist}_k(o)\}$$

Note: $|N_k(o)|$ may contain more than k objects, because objects may have same distance.

Let o' be a neighbour of o , to avoid $\text{dist}(o, o')$ is too small, we compute the reachability distance from o to o' :

$$\text{reachdist}_k(o' \leftarrow o) = \max\{\text{dist}_k(o), \text{dist}(o, o')\}$$

Note that reachability distance is not symmetric, thus

$$\text{reachdist}_k(o \leftarrow o') \neq \text{reachdist}_k(o' \leftarrow o)$$

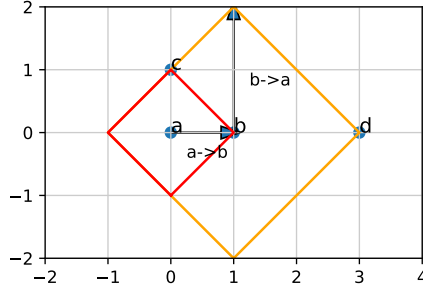


Figure 4: $\text{reachdist}_2(b \leftarrow a) \neq \text{reachdist}_2(a \leftarrow b)$

$\text{reachdist}_2(b \leftarrow a)$	$= \max\{\text{dist}_2(a), \text{dist}(a, b)\}$	$= \max\{1, 1\}$	$= 1$
$\text{reachdist}_2(c \leftarrow a)$	$= \max\{\text{dist}_2(a), \text{dist}(a, c)\}$	$= \max\{1, 1\}$	$= 1$
$\text{reachdist}_2(d \leftarrow a)$	$= \max\{\text{dist}_2(a), \text{dist}(a, d)\}$	$= \max\{1, 3\}$	$= 3$
$\text{reachdist}_2(a \leftarrow b)$	$= \max\{\text{dist}_2(b), \text{dist}(b, a)\}$	$= \max\{2, 1\}$	$= 2$
$\text{reachdist}_2(c \leftarrow b)$	$= \max\{\text{dist}_2(b), \text{dist}(b, c)\}$	$= \max\{2, 2\}$	$= 2$
$\text{reachdist}_2(d \leftarrow b)$	$= \max\{\text{dist}_2(b), \text{dist}(b, d)\}$	$= \max\{2, 2\}$	$= 2$
$\text{reachdist}_2(a \leftarrow c)$	$= \max\{\text{dist}_2(c), \text{dist}(c, a)\}$	$= \max\{2, 1\}$	$= 2$
$\text{reachdist}_2(b \leftarrow c)$	$= \max\{\text{dist}_2(c), \text{dist}(c, b)\}$	$= \max\{2, 2\}$	$= 2$
$\text{reachdist}_2(d \leftarrow c)$	$= \max\{\text{dist}_2(c), \text{dist}(c, d)\}$	$= \max\{2, 4\}$	$= 4$
$\text{reachdist}_2(a \leftarrow d)$	$= \max\{\text{dist}_2(d), \text{dist}(d, a)\}$	$= \max\{3, 3\}$	$= 3$
$\text{reachdist}_2(b \leftarrow d)$	$= \max\{\text{dist}_2(d), \text{dist}(d, b)\}$	$= \max\{3, 2\}$	$= 3$
$\text{reachdist}_2(c \leftarrow d)$	$= \max\{\text{dist}_2(d), \text{dist}(d, c)\}$	$= \max\{3, 4\}$	$= 4$

Note that,

$$\text{reachdist}_k(o \leftarrow o') \neq \text{reachdist}_k(o' \leftarrow o)$$

The *Local Reachability Density* (LRD) of an object o is defined as

$$\text{lrd}_k(o) = \frac{|N_k(o)|}{\sum_{o' \in N_k(o)} \text{reachdist}_k(o \leftarrow o')}$$

$$\begin{aligned}
\text{lrd}_2(a) &= |N_2(a)| / (\text{reachdist}_2(a \leftarrow b) + \text{reachdist}_2(a \leftarrow c)) &= 2 / (2 + 2) &= 0.5 \\
\text{lrd}_2(b) &= |N_2(b)| / (\text{reachdist}_2(b \leftarrow a) + \text{reachdist}_2(b \leftarrow c) + \text{reachdist}_2(b \leftarrow d)) &= 3 / (1 + 2 + 3) &= 0.5 \\
\text{lrd}_2(c) &= |N_2(c)| / (\text{reachdist}_2(c \leftarrow a) + \text{reachdist}_2(c \leftarrow b)) &= 2 / (1 + 2) &= 0.667 \\
\text{lrd}_2(d) &= |N_2(d)| / (\text{reachdist}_2(d \leftarrow a) + \text{reachdist}_2(d \leftarrow b)) &= 2 / (3 + 2) &= 0.4
\end{aligned}$$

The *Local Outlier Factor* (LOF) of an object o is

$$\text{LOF}_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{\text{lrd}_k(o')}{\text{lrd}_k(o)}}{|N_k(o)|} = \frac{\sum_{o' \in N_k(o)} \text{lrd}_k(o')}{|N_k(o)| \cdot \text{lrd}_k(o)}$$

$$\begin{aligned}
\text{LOF}_2(a) &= (\text{lrd}_2(b) + \text{lrd}_2(c)) / (|N_2(a)| \cdot \text{lrd}_2(a)) &= (0.5 + 0.667) / (2 \times 0.5) &= 1.167 \\
\text{LOF}_2(b) &= (\text{lrd}_2(a) + \text{lrd}_2(c) + \text{lrd}_2(d)) / (|N_2(b)| \cdot \text{lrd}_2(b)) &= (0.5 + 0.667 + 0.4) / (3 \times 0.5) &\approx 1.045 \\
\text{LOF}_2(c) &= (\text{lrd}_2(a) + \text{lrd}_2(b)) / (|N_2(c)| \cdot \text{lrd}_2(c)) &= (0.5 + 0.5) / (2 \times 0.667) &\approx 0.750 \\
\text{LOF}_2(d) &= (\text{lrd}_2(a) + \text{lrd}_2(b)) / (|N_2(d)| \cdot \text{lrd}_2(d)) &= (0.5 + 0.5) / (2 \times 0.4) &= 1.25
\end{aligned}$$

LOF is a relative value. We should not compare a LOF from one dataset with other reading from different datasets.

The question does not ask for outlier. There is only 4 data points. If we remove 1 point, we will remove 25% of the data. I do not determine the outlier here.