Tut 10: Hierarchical Clustering

May/June 2022

Hierarchical Clustering

1. Suppose that we have four observations, for which we compute a distance matrix:

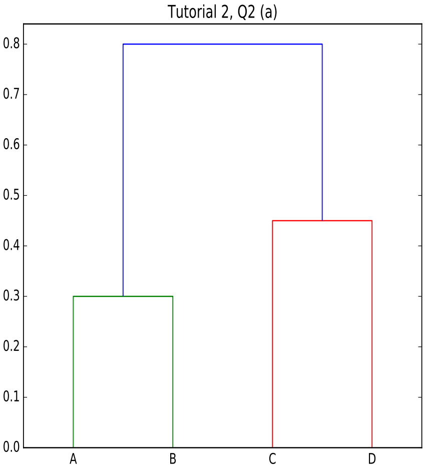
$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

   (a) Sketch the dendrogram that results from hierarchically clustering these four observations using **complete linkage**. Plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram. Suppose that we cut the dendrogram such that two clusters result. What are the observations in each cluster?

Solution. Use formula for complete linkage.

Step 1: $d(A, B) = 0.3 \Rightarrow$ merge A,B

$$\begin{array}{c|cccc|} & A & B & C & D \\ \hline A & 0 & & & \\ B & \boxed{0.3} & 0 & & \\ C & 0.4 & 0.5 & 0 & \\ D & 0.7 & 0.8 & 0.45 & 0 \\ \hline \end{array}$$



Step 2: $d(C, D) = 0.45 \Rightarrow$ merge C,D

$$\begin{array}{c|ccc|} & AB & C & D \\ \hline AB & 0 & & \\ C & 0.5 & 0 & \\ D & 0.8 & \boxed{0.45} & 0 \\ \hline \end{array}$$

Step 3:

$$\begin{array}{c|cc|} & AB & CD \\ \hline AB & 0 & \\ CD & \boxed{0.8} & 0 \\ \hline \end{array}$$

```
import numpy as np, matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial.distance import squareform

# https://stackoverflow.com/questions/41416498/dendrogram-or-other-plot-
mat = np.array([[0.0, 0.3, 0.4, 0.7], [0.3, 0.0, 0.5, 0.8],
  [0.4, 0.5, 0.0, 0.45], [0.7, 0.8, 0.45, 0.0]])
dists = squareform(mat)
```

```
linkage_matrix = linkage(dists, "complete")
dendrogram(linkage_matrix, labels=list("ABCD"))
plt.title("Tutorial 2, Q2 (a)")
plt.show()
```

□

(b) Repeat (a) using single linkage clustering.

Solution. Use single-linkage formula.

Step 1: $d(A, B) = 0.3 \Rightarrow$ merge A,B

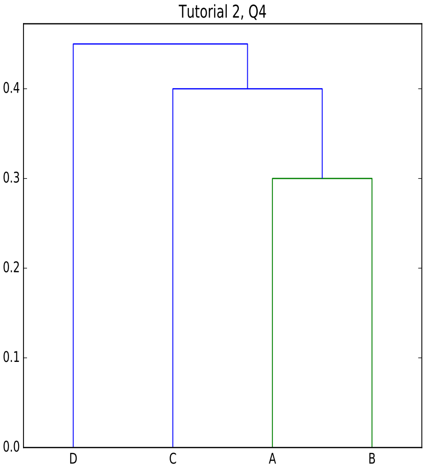
$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & & & \\ B & \boxed{0.3} & 0 & & \\ C & 0.4 & 0.5 & 0 & \\ D & 0.7 & 0.8 & 0.45 & 0 \end{array}$$

Step 2: $d((A, B), C) = 0.4 \Rightarrow$ merge (A,B), C

$$\begin{array}{c|ccc} & AB & C & D \\ \hline AB & 0 & & \\ C & \boxed{0.4} & 0 & \\ D & 0.7 & 0.45 & 0 \end{array}$$

Step 3:

$$\begin{array}{c|cc} & ABC & D \\ \hline ABC & 0 & \\ D & \boxed{0.45} & 0 \end{array}$$



□

2. (May 2020 Final Q3(a)) Given the unlabelled data in Table 3.1.

Table 3.1: Unlabelled data.

	V1	V2	V3
1	7.5205	4.6564	-0.1947
2	-1.1824	-1.1174	1.8383
3	-0.3576	-0.4739	-1.1603
4	-1.422	-0.5891	-0.8287
5	3.2287	0.7141	0.6208
6	3.2926	3.1609	2.7553
7	8.2304	3.8832	-1.7378
8	4.2079	0.4964	4.361
9	3.8443	5.7565	1.0293
10	1.493	3.525	-2.9904

Use the k-means algorithm with $k = 2$ (unsupervised learning) to find the final cluster centres if the **first** and **sixth** rows are chosen as the **initial cluster centres**. (4 marks)

Solution. Given the initial centres:

V1	V2	V3
7.5205	4.6564	-0.1947
3.2926	3.1609	2.7553

Step 1 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	0	5.3679	A
-1.1824	-1.1174	1.8383	10.64	6.2586	B
-0.3576	-0.4739	-1.1603	9.4508	6.4705	B
-1.422	-0.5891	-0.8287	10.3868	7.0096	B
3.2287	0.7141	0.6208	5.8844	3.2476	B
3.2926	3.1609	2.7553	5.3679	0	B
8.2304	3.8832	-1.7378	1.8663	6.715	A
4.2079	0.4964	4.361	7.0024	3.2428	B
3.8443	5.7565	1.0293	4.0278	3.1655	B
1.493	3.525	-2.9904	6.7399	6.0319	B

. . . . [1.5 marks]

The new cluster centres are

V1	V2	V3
7.87545	4.2698	-0.96625
1.6380625	1.4340625	0.7031625

. . . . . . [0.5 mark]

Step 2 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	0.9331	6.767	A
-1.1824	-1.1174	1.8383	10.9056	3.9691	B
-0.3576	-0.4739	-1.1603	9.5039	3.331	B
-1.422	-0.5891	-0.8287	10.4914	3.9754	B
3.2287	0.7141	0.6208	6.0625	1.7479	B
3.2926	3.1609	2.7553	6.0068	3.1513	B
8.2304	3.8832	-1.7378	0.9331	7.4442	A
4.2079	0.4964	4.361	7.4879	4.5676	B
3.8443	5.7565	1.0293	4.7374	4.8639	A
1.493	3.525	-2.9904	6.737	4.2468	B

. . . . [0.5 mark]

The new cluster centres are

V1	V2	V3
6.53173333333333	4.76536666666667	-0.301066666666667
1.32288571428571	0.816571428571429	0.656571428571428

. . . . [0.5 mark]

Step 3 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	1.0004	7.3403	A
-1.1824	-1.1174	1.8383	9.9344	3.3783	B
-0.3576	-0.4739	-1.1603	8.6978	2.7911	B
-1.422	-0.5891	-0.8287	9.6026	3.4229	B
3.2287	0.7141	0.6208	5.3078	1.9089	B
3.2926	3.1609	2.7553	4.7337	3.7122	B
8.2304	3.8832	-1.7378	2.3933	7.9279	A
4.2079	0.4964	4.361	6.7349	4.7062	B
3.8443	5.7565	1.0293	3.1582	5.5587	A
1.493	3.525	-2.9904	5.8446	4.5459	B

. . . . [0.5 mark]

There is no change in the clustering, the final cluster centres are

V1	V2	V3
6.53173333333333	4.76536666666667	-0.301066666666667
1.32288571428571	0.816571428571429	0.656571428571428

. . . . [0.5 mark]

□

3. (Jan 2022 Final Q3) Given the two-dimensional data in Table 3.1.

Table 3.1: Two-dimensional data for clustering.

	x_1	x_2
A	2	0
B	3	1
C	4	3
D	0.5	1
E	1	2.5
F	2.5	3.3

(a) Perform k-means clustering algorithm (using the Euclidean algorithm) on the data from Table 3.3 with A and B as the initial centres until two clusters are found. Write down the stable cluster centres. You may round the numbers in your calculations to 4 decimal places. (13 marks)

Solution. Step 1 : Update table based on distance to cluster centres

x_1	x_2	dist.1	dist.2	clust.centre
2	0	0	1.4142	①
3	1	1.4142	0	②
4	3	3.6056	2.2361	②
0.5	1	1.8028	2.5	①
1	2.5	2.6926	2.5	②
2.5	3.3	3.3377	2.3537	②

. . . . [7 marks]

The new cluster centres are

$$\begin{aligned}\text{Cluster 1 centre} &= \frac{(2,0)+(0.5,1)}{2} = (1.25, 0.5)\\ \text{Cluster 2 centre} &= \frac{(3,1)+(4,3)+(1,2.5)+(2.5,3.3)}{4} = (2.625, 2.45)\end{aligned}$$

[1 mark]

Step 2 : Update table based on distance to the new cluster centres

x_1	x_2	dist.1	dist.2	clust.centre
2	0	0.9014	2.5285	①
3	1	1.82	1.4977	②
4	3	3.7165	1.4809	②
0.5	1	0.9014	2.5726	①
1	2.5	2.0156	1.6258	②
2.5	3.3	3.0663	0.8591	②

. . . . [4 marks]

The new cluster centres remain the same as the previous step, the k-means algorithm stops and have

$$(1.25, 0.5), \quad (2.625, 2.45)$$

as the stable cluster centres. . . . . [1 mark] □

(b) Construct the hierarchical clustering with single linkage for the data in Table 3.3. Suppose the distance table for the points A to E is obtained as follows:

	A	B	C	D	E
A	0				
B	1.4142	0			
C	3.6056	2.2361	0		
D	1.8028	2.5000	4.0311	0	
E	2.6926	2.5000	3.0414	1.5811	0

Expand the distance to the data in Table 3.3 to all the points A to F and then perform the necessary steps (you may want to write your answer in pencil because it is easy to get the updated distance matrices wrong) to draw the dendrogram with proper labels. (10 marks)

Solution. The distance from point A to E against F are given below:

$$\begin{aligned}dist(A,F) &= \sqrt{(2-2.5)^2+(0-3.3)^2} = 3.3377\\ dist(B,F) &= \sqrt{(3-2.5)^2+(1-3.3)^2} = 2.3537\\ dist(C,F) &= \sqrt{(4-2.5)^2+(3-3.3)^2} = 1.5297\\ dist(D,F) &= \sqrt{(0.5-2.5)^2+(1-3.3)^2} = 3.0480\\ dist(E,F) &= \sqrt{(1-2.5)^2+(2.5-3.3)^2} = 1.7\end{aligned}$$

The distance table to all the points A to F is

	A	B	C	D	E	F

```
A            0
B      1.4142      0
C      3.6056 2.2361      0
D      1.8028 2.5000 4.0311      0
E      2.6926 2.5000 3.0414 1.5811      0
F      3.3377 2.3537 1.5297 3.0480 1.7000      0
```

. . . . [3.5 marks]

The minimum distance is 1.4142, so A and B should be grouped. . . . . . . [0.5 mark]

```
           AB      C      D      E
AB          0
C      2.2361      0
D      1.8028 4.0311      0
E      2.5000 3.0414 1.5811      0
F      2.3537 1.5297 3.0480 1.7000
```

. . . . [2 marks]

The minimum distance is 1.5297, so C and F should be grouped.

```
           AB     CF      D      E
AB          0
CF     2.2361      0
D      1.8028 3.0480      0
E      2.5000 1.7000 1.5811      0
```

. . . . [1 mark]

The minimum distance is 1.5811, so D and E should be grouped.

```
           AB     CF     DE
AB          0
CF     2.2361      0
DE     1.8028 1.7000      0
```

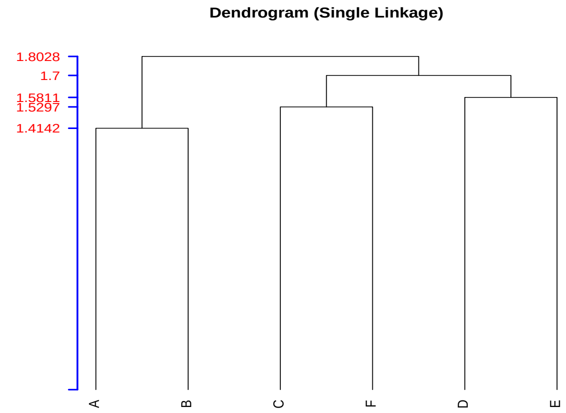
. . . . [1 mark]

The minimum distance is 1.7000, so CF and DE should be grouped.

```
           AB  CF,DE
AB          0
CF,DE  1.8028      0
```

. . . . [1 mark]

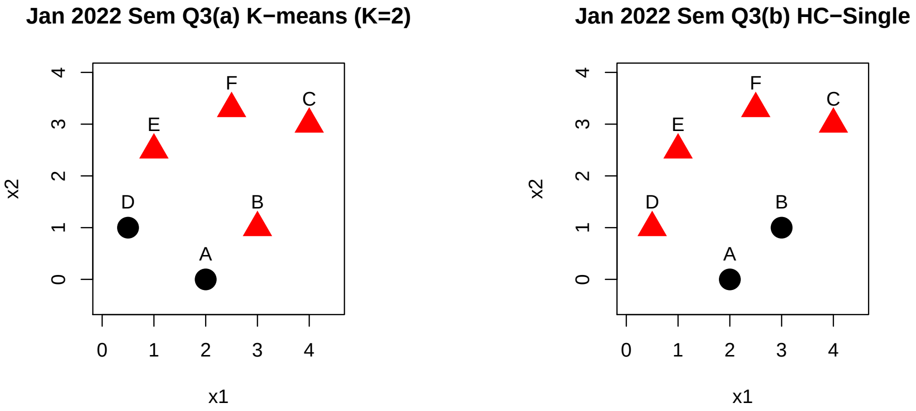
We can now sketch the dendrogram:



. . . . [1 mark] □

(c) Sketch (with appropriate labels) the clusters obtained from the k-means clustering in part (a) and the clusters obtained from hierarchical clustering with single linkage by cutting the dendrogram into two subtrees. (2 marks)

Solution. The sketch of k-means clustering is on the left while the sketch of the two clusters from the hierarchical clustering with single linkage is on the right.



With appropriate labels . . . . [2 marks] □