Tut 10: Hierarchical Clustering

June 2024

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

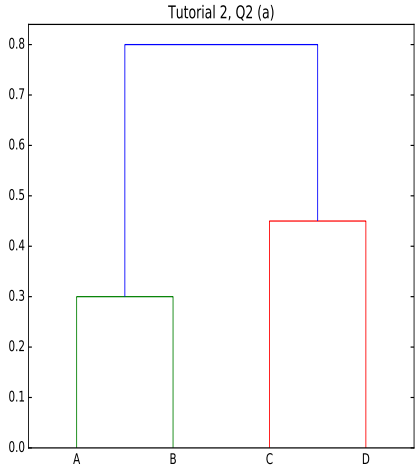
	A	B	C	D
A	0			
B	0.3	0		
C	0.4	0.5	0	
D	0.7	0.8	0.45	0

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

	AB	C	D
AB	0		
C	0.5	0	
D	0.8	0.45	0

Step 3:

	AB	CD
AB	0	
CD	0.8	0




```
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-from-
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

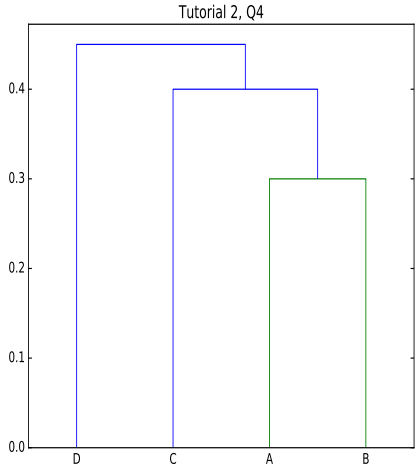
	A	B	C	D
A	0			
B	0.3	0		
C	0.4	0.5	0	
D	0.7	0.8	0.45	0

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

	AB	C	D
AB	0		
C	0.4	0	
D	0.7	0.45	0

Step 3:

	ABC	D
ABC	0	
D	0.45	0



□

(c) Repeat (a) using average linkage clustering. (Exercise)

2. (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 distance) on the data from Table 3.1 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

$$\text{Cluster 1 centre} = \frac{(2,0)+(0.5,1)}{2} = (1.25, 0.5)$$

$$\text{Cluster 2 centre} = \frac{(3,1)+(4,3)+(1,2.5)+(2.5,3.3)}{4} = (2.625, 2.45)$$

[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 3.1. 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 matrix to the data in Table 3.1 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:

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

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

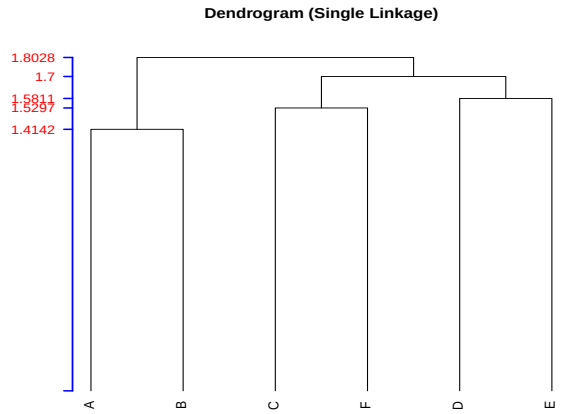
.....[1 mark]

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]

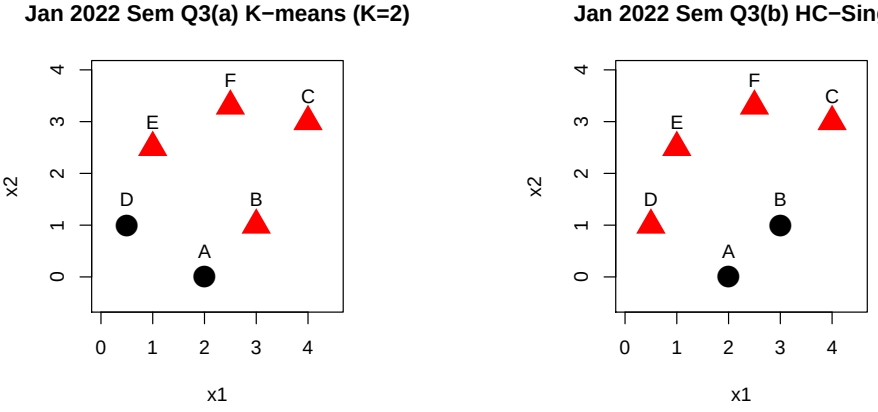
We can now sketch the dendrogram:



.....[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] □

3. (Final Exam Jan 2023, Q5(b)) Given the unlabelled data in Table 5.2.

Obs.	x_1	x_2
A	0.28	0.13
B	0.12	0.31
C	0.21	3.51
D	0.62	3.62
E	4.15	4.32
F	4.52	4.27
G	4.34	3.89

Table 5.2: Unlabelled data.

(i) Suppose the (Euclidean) distance matrix of Table 5.2 is given below:

	A	B	C	D	E	F	G
A	0						
B	0.241	0					
C	3.381	3.201	0				
D	3.507	3.348	0.424	0			
E	5.704	5.685	4.022	3.599	0		
F	5.926	5.920	4.376	3.954	0.373	0	
G	5.534	5.534	4.147	3.730	0.470	0.420	0

Construct the hierarchical clustering with **Euclidean distance** and **single linkage** and then draw the **dendrogram** of the hierarchical clustering. (9 marks)

Solution. Height = 0.241; Cluster: A, B

	A,B	C	D	E	F	G
A,B	0					
C	3.201	0				
D	3.348	0.424	0			
E	5.685	4.022	3.599	0		
F	5.920	4.376	3.954	0.373	0	
G	5.534	4.147	3.730	0.470	0.420	0

. . . . . [2 marks]

Height = 0.373; Cluster: E, F

	A,B	C	D	E,F	G
A,B	0				
C	3.201	0			
D	3.348	0.424	0		
E,F	5.685	4.022	3.599	0	
G	5.534	4.147	3.730	0.420	0

. . . . . [2 marks]

Height = 0.420; Cluster: (E,F), G

	A,B	C	D	EF,G
A,B	0			
C	3.201	0		
D	3.348	0.424	0	
EF,G	5.534	4.022	3.599	0

. . . . . [1 mark]

Height = 0.424; Cluster: C, D

	A,B	C,D	EF,G
A,B	0		
C,D	3.201	0	
EF,G	5.534	3.599	0

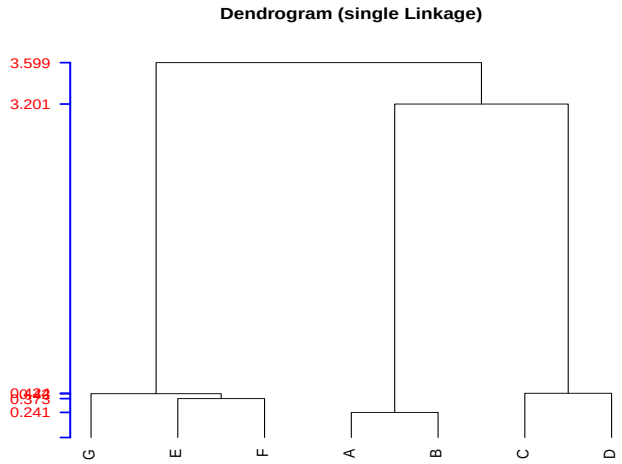
. . . . . [1 mark]

Height = 3.201; Cluster: CD, AB

	AB,CD	EF,G
AB,CD	0	
EF,G	3.599	0

. . . . . [1 mark]

The dendrogram is shown below:



Marks are deducted for poor labelling or terribly drawn lines . . . . . [2 marks]
Average: 6.60 / 9 marks in Jan 2023; 15% below 4.5 marks.

□

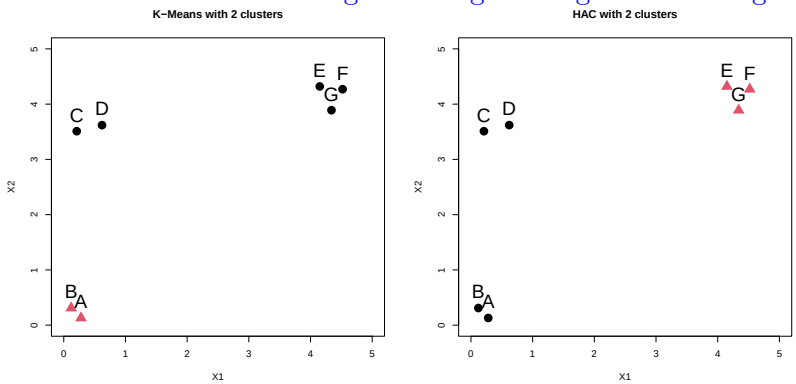
(ii) Sketch (with **appropriate symbols/labels**) the clusters obtained from the k-means clustering (with $k = 2$) below:

```
Cluster means:
     X1    X2
1 2.768 3.922
2 0.200 0.220

Clustering vector:
A B C D E F G
2 2 1 1 1 1 1
```

and the hierarchical clustering in part (i) by cutting the dendrogram into two subtrees.
(3 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. . . . . [3 marks]
Average: 0.61 / 3 marks in Jan 2023; 54% below 1.5 marks.

□

4. (Final Exam Jan 2024 Sem, Q3) Given the three-dimensional data in Table 3.1.

Table 3.1: Three-dimensional data

Obs.	x_1	x_2	x_3
A	4	5	3
B	7	4	3
C	4	10	2
D	0	2	6
E	4	7	1
F	1	3	4

(a) Write down the mathematical formula of the Minkowski distance of order $r(\geq 1)$ for two vectors (x_1, x_2, x_3) and (y_1, y_2, y_3). (2 marks)

Solution. $\left(|x_1 - y_1|^r + |x_2 - y_2|^r + |x_3 - y_3|^r\right)^{1/r}$, $r \geq 1$. .....[2 marks]
Average: 1.11 / 2 marks in Jan 2024; 38.18% below 1 mark. □

(b) Suppose a partial (Euclidean) distance matrix of Table 3.1 is given below:

```
        A        B        C        D
A   0
B   3.1623
C   5.0990   6.7823
D   5.8310   7.8740   9.7980
E   2.8284   4.6904   3.1623   8.1240
F   3.7417   6.1644   7.8740   2.4495
```

i. Calculate the Euclidean distance of the point E to the point F and write down the complete distance matrix for the three-dimensional data in Table 3.1. (3 marks)

Solution. $\left(|4-1|^2 + |7-3|^2 + |1-4|^2\right)^{1/2} = \sqrt{9+16+9} = 5.8310$ ...[2 marks]
The complete distance matrix for the three-dimensional data in Table 3.1 is

```
          A        B        C        D        E        F
A         0
B   3.1623         0
C   5.0990   6.7823         0
D   5.8310   7.8740   9.7980         0
E   2.8284   4.6904   3.1623   8.1240         0
F   3.7417   6.1644   7.8740   2.4495   5.8310         0
```

.....[1 mark]
Average: 2.69 / 3 marks in Jan 2024; 9.09% below 1.5 marks. □

ii. Use the complete distance matrix in part (i) to perform hierarchical clustering analysis with **complete linkage** and then draw the **dendrogram** of the hierarchical clustering. (10 marks)

Solution. The first height is 2.4495, D and F are merged as follows: ....[1 mark]

```
          A        B        C      D,F        E
A         0
B   3.1623         0
C   5.0990   6.7823         0
D,F 5.8310   7.8740   9.7980         0
E   2.8284   4.6904   3.1623   8.1240         0
```

.....[2 marks]
The second height is 2.8284, A and E are merged as follows: .....[1 mark]

```
        A,E        B        C      D,F
A,E       0
B   4.6904         0
C   5.0990   6.7823         0
D,F 8.1240   7.8740   9.7980         0
```

..... [1 mark]

The third height is 4.6904, AE and B are merged as follows: ..... [1 mark]

```
        AE,B        C        D,F
AE,B       0
C     6.7823
D,F   8.1240   9.7980        0
```

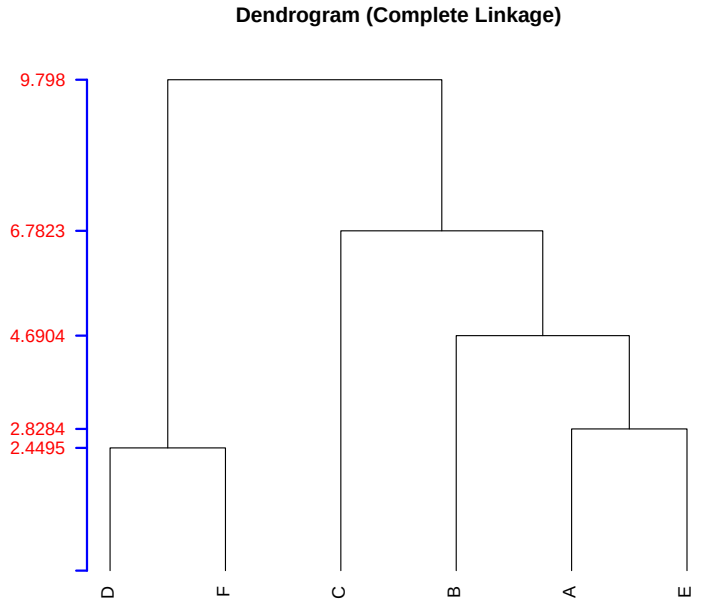
..... [1 mark]

The fourth height is 6.7823, AEB and C are merged as follows:

```
        AEB,C       D,F
AEB,C       0
D,F    9.7980         0
```

..... [1 mark]

The dendrogram is drawn below:



Marks are deducted for poor labelling or terribly drawn lines (i.e. there should not have terrible crossings) ..... [2 marks]

Average: 7.75 / 10 marks in Jan 2024; 14.55% below 5 marks. □

(c) Perform k-means clustering algorithm using the Euclidean distance on the data from Table 3.1 with B and C 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.

(6 marks)

Solution. Given the initial centres B(7, 4, 3), C(4, 10, 2) which correspond to cluster 1 and cluster 2.

Step 1 : Update table based on distance to cluster centres (the distance can be obtained from part (b)(i))

x_1	x_2	x_3	dist.1	dist.2	clust.centre
4	5	3	3.1623	5.0990	1
7	4	3	0	6.7823	1
4	10	2	6.7823	0	2
0	2	6	7.8740	9.7980	1
4	7	1	4.6904	3.1623	2
1	3	4	6.1644	7.8740	1

..... [2 marks]

The new cluster centres are

$$Centre1 = (3, 3.5, 4), \quad Centre2 = (4, 8.5, 1.5). \qquad \text{[1 mark]}$$

Step 2 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
4	5	3	2.0616	3.8079	1
7	4	3	4.1533	5.6125	1
4	10	2	6.8739	1.5811	2
0	2	6	3.9051	8.8600	1
4	7	1	4.7170	1.5811	2
1	3	4	2.0616	6.7454	1

.....[2 marks]

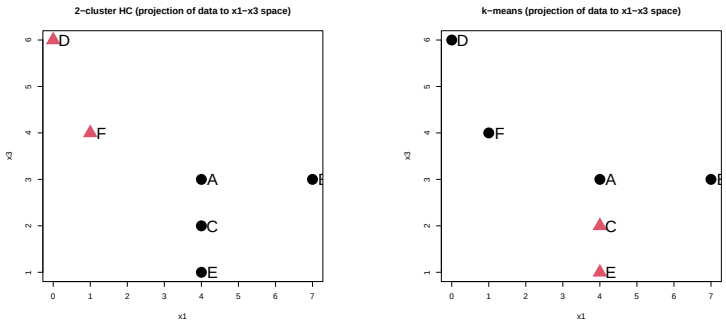
The stable cluster centres are

$$Centre1 = (3, 3.5, 4), \quad Centre2 = (4, 8.5, 1.5). \qquad [1 \text{ mark}]$$

Average: 5.07 / 6 marks in Jan 2024; 12.73% below 3 marks. □

(d) With **appropriate symbols/labels**, sketch the projections of the data in Table 3.1 to the subspace spanned by the variables x_1 and x_3 with the cluster labels from the hierarchical clustering analysis from part (b) and the cluster labels from the k-means clustering from part (c). (4 marks)

Solution. The hierarchical clustering is shown on the left and the k-means is shown on the right. ..... [2 × 2 = 4 marks]



Average: 0.87 / 4 marks in Jan 2024; 76.36% below 2 marks. □