Tut 10: Hierarchical Clustering

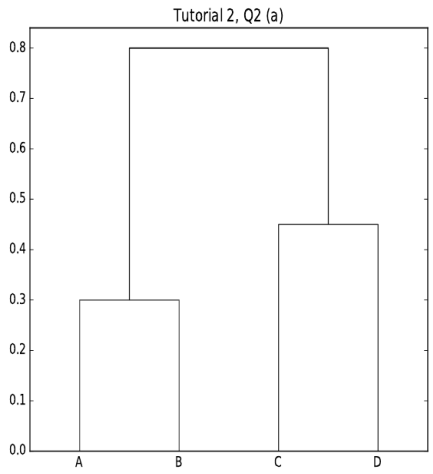
May/June 2022

Hierarchical Clustering

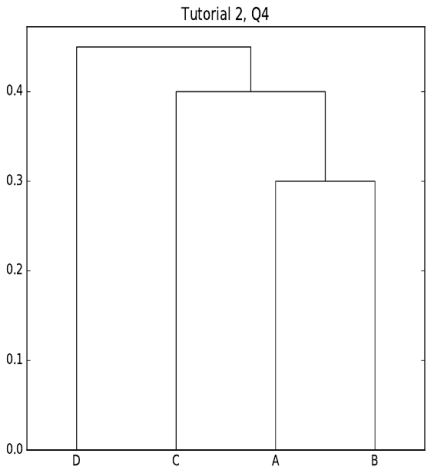
1. Suppose that we have four observations, for which we compute a distance matrix:

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

   (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?



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

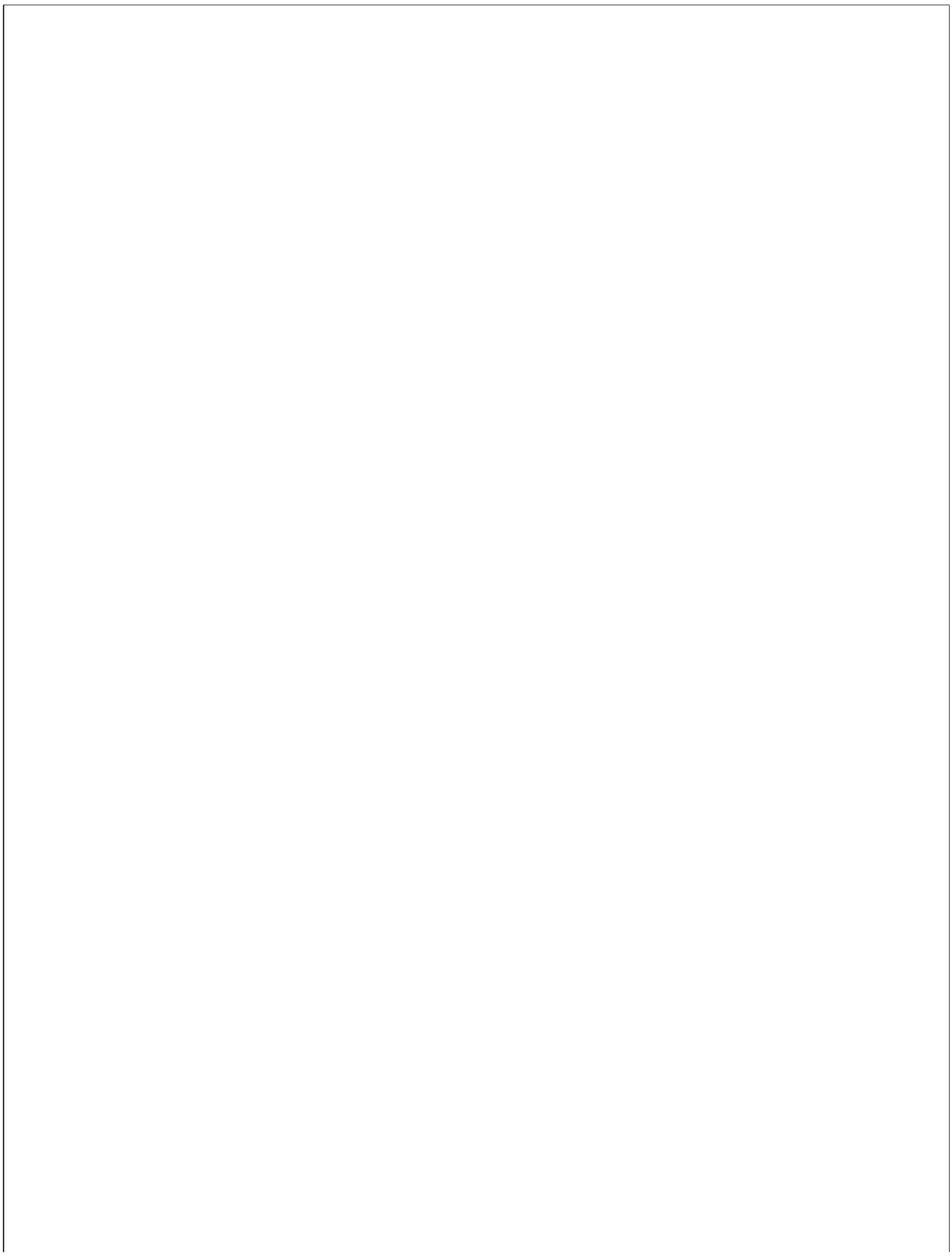


2. (May 2020 Final Q3(a)) Given the unlabelled data in Table 3.1.

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)



3. (Jan 2020 Final Q3) Given the two-dimensional data in Table 3.1.

Table 3.1: Two-dimensional data for clustering.

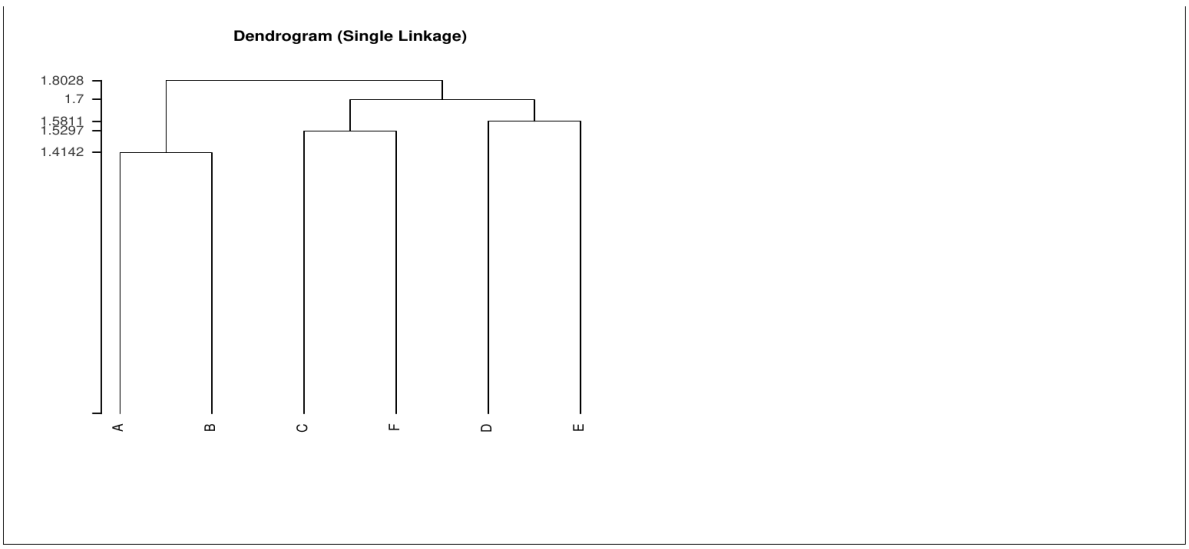
	x_1	x_2
A	2	0
B	3	1
C	4	3
D	0.5	1
E	1	2.5
F	2.5	3.3

(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)

(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)



(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)