Data Mining
Assignment A2

Sindhuja Yerramalla
U00839259

Q1

Given 8 two-dimensional points

x_1 (15,10) x_2 (3,10) x_3 (15,12) x_4 (3,14) x_5 (18,13)

x_6 (1,7) x_7 (10,1) x_8 (10,30)

(i) If K=2, given initial means (10,1) & (10,30)

Euclidean distance = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

iteration 1 :

	point	(10,1) distmean1	(10,30) distmean2	cluster
x_1	(15,10)	10.29	20.61	1
x_2	(3,10)	11.40	21.18	1
x_3	(15,12)	12.08	18.68	1
x_4	(3,14)	14.76	17.46	1
x_5	(18,13)	14.42	18.78	1
x_6	(1,7)	10.81	24.69	1
x_7	(10,1)	0	29	1
x_8	(10,30)	29	0	2

after 1st iteration we have

cluster1 :- x_1 (15,10) x_2 (3,10) x_3 (15,12) x_4 (3,14)
x_5 (18,13) x_6 (1,7) x_7 (10,1)

cluster2 :- x_8 (10,30)

Centroid $= \left[\frac{\sum x_i}{n}, \frac{\sum y_i}{n}\right]$

Centroid of cluster 1 $= \left[\frac{15+3+15+3+18+1+10}{7}, \frac{10+10+12+14+13+7+1}{7}\right]$

$K_1 = (9.28, 9.57)$

Centroid of cluster 2 $K_2 = (10, 30)$

2nd iteration :

		(9.28, 9.57) dist mean1	(10, 30) dist mean2	cluster
x_1	(15,10)	5.73	20.61	1
x_2	(3,10)	6.29	21.18	1
x_3	(15,12)	6.21	18.68	1
x_4	(3,14)	7.68	17.46	1
x_5	(18,13)	9.37	18.78	1
x_6	(1,7)	8.66	24.69	1
x_7	(10,1)	8.60	29	1
x_6	(10,30)	20.44	0	2.

The K-mean clustering ends when our clustering method from first iteration to other iteration does not change.

∴ Final Allocation after 2 iteration, when K=2 is

- $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ belongs to cluster1 with $K_1(9.28, 9.57)$ as centroid
- x_8 belongs to cluster 2 with $K_2(10,30)$ as centroid

(11) If K=3. Given initial means as (10,1) (10,30) (3,10)

	Point	(10,1) distmeall	(10,30) distmean2	(3,10) distmean3	cluster
x_1	(15,10)	10.29.	20.61	12.	1
x_2	(3,10)	11.40	21.18	0	3
x_3	(15,12)	12.08	18.68	12.16	1
x_4	(3,14)	14.76	17.46	4	3
x_5	(18,13)	14.42.	18.78	15.29.	1
x_6	(1,7)	10.81	24.69	3.60	3
x_7	(10,1)	0	29	11.40	1
x_8	(10,30)	29.	0	21.18	2.

After 1st iteration we have

cluster1 :- $x_1(15,10)$, $x_3(15,12)$, $x_5(18,13)$, $x_7(10,1)$

cluster2 :- $x_8(10,30)$

cluster3 :- $x_2(3,10)$, $x_4(3,14)$, $x_6(1,7)$

∴ centroids $K_1(14.5, 9)$, $K_2(10,30)$, $K_3(2.33, 10.33)$

Iteration 2

		(14.5,9)	(10,30)	(2.33, 10.33)	Cluster
x_1	(15,10)	1.11	20.61	12.67	1
x_2	(3,10)	11.54	21.18	0.74	3
x_3	(15,12)	3.04	18.68	12.78	1
x_4	(3,14)	12.53	17.46	3.73	3
x_5	(18,13)	5.31	18.78	15.90	1
x_6	(1,7)	13.64	24.69	3.59	3
x_7	(10,1)	9.17	29.	12.07	1
x_8	(10,30)	21.47	0.	21.11	2

Final Allocations after 2nd iteration are when k=3 are.

- x_1 (15,10), x_3 (15,12), x_5 (18,13), x_7 (10,1) belongs to cluster 1 with centroid K_1 (14.5, 9).
- x_8 (10,30) belongs to cluster 2 with centroid K_2 K_2 (10, 30)
- x_2 (3,10), x_4 (3,14), x_6 (1,7) belongs to cluster 3 with centroid K_3 (2.33, 10.33).

Q2 Advantages of K-means algorithm

* Easy to compute and implement, since the only task is to calculate Euclidean distance.
* Every point in the dataset will be alloted to a cluster. Therefore there will be no ambiguity.
* Solves 1 dimensional problems in polynomial time.
* This algorithm converges very fast.

Disadvantages of K-means algorithm

* The complexity of problem increases as the number of dimensions increase.
* The output will be inaccurate if K and value and intial centroids are not picked carectly
* Doesn't precompute clusters before solving the problem
* This will have effeciency problems because we have to evaluate euclidean distance for each point in every iteration

Q3

	a	b	c	d	e	f	g	h
a	0							
b	11	0						
c	5	13	0					
d	12	2	14	0				
e	7	17	1	18	0			
f	13	4	15	5	20	0		
g	9	15	12	16	15	19	0	
h	11	20	12	21	17	22	30	0

As distance '1' between 'c' & 'e' is minimum, we will merge them in first iteration.

	a	b	c,e	d	f	g	h
a	0						
b	11	0					
c,e	6	15	0				
d	12	2	16	0			
f	13	17	17.5	5	0		
g	9	15	13.5	16	19	0	
h	11	20	14.5	21	22	30	0

The next minimum distance is '2' and it is there between 'b' and 'd', we will merge them and get the average linkage between merged and other data points

	a	b,d	c,e	f	g	h
a	0					
b,d	11.5	0				
c,e	6	15.5	0			
f	13	4.5	17.5	0		
g	9	15.5	13.5	19	0	
h	11	20.5	14.5	22	30	0

The next minimum distance is 4.5, it is between 'f' and 'b,d', we will next merge these two by doing average between the points

	a	b,d,f	c,e	g	h
a	0				
b,d,f	12.25	0			
c,e	6	16.5	0		
g	9	17.25	13.5	0	
h	11	21.25	14.5	30	0

Now, minimum distance of 6 exists between 'c,e' and 'a', we will put them in same cluster

	a,c,e	b,d,f	g	h
a,c,e	0			
b,d,f	14.375	0		
g	11.25	17.25	0	
h	12.75	21.25	30	0

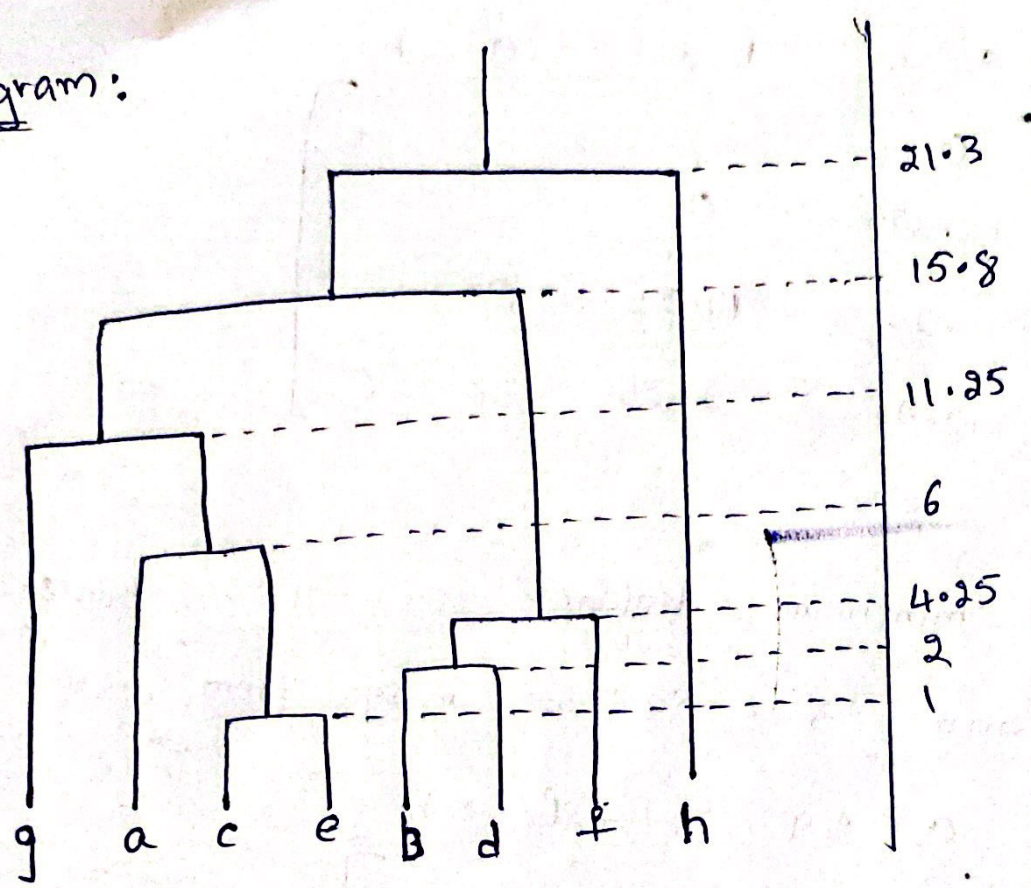
The next minimum distance is '11.25', between 'a,c,e' and 'g'. so we merge them now

	a,c,e,g	b,d,f	h
a,c,e,g	0		
b,d,f	15.8	0	
h.	21.375	21.25	0

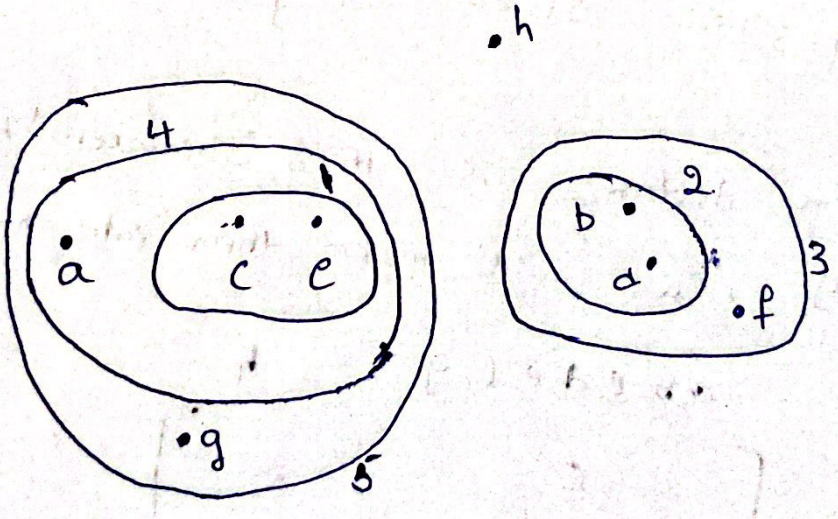
Now, minimum distance is 15.8 between 'b,d,f' and a,c,e,g. we will make them into one cluster

	a,b,c,d,e,f,g	h
a,b,c,d,e,f,g	0	
h	21.312	0

Dendogram:



5 clusters



1st cluster - c,e

2nd cluster - b,d

3rd cluster - b, d, f

4th cluster - a, c, e.

5th cluster - a, c, e, g