

Tutorial exercises

Clustering – K-means, Nearest Neighbor and Hierarchical.

Exercise 1. K-means clustering

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters:

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

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)
- The centers of the new clusters
- Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
- How many more iterations are needed to converge? Draw the result for each epoch.

Solution:

a)

$d(a,b)$ denotes the Euclidean distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b) = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:

$d(A1, \text{seed1})=0$ as A1 is seed1

$d(A1, \text{seed2})= \sqrt{13} > 0$

$d(A1, \text{seed3})= \sqrt{65} > 0$

→ A1 ∈ cluster1

A2:

$d(A2, \text{seed1})= \sqrt{25} = 5$

$d(A2, \text{seed2})= \sqrt{18} = 4.24$

$d(A2, \text{seed3})= \sqrt{10} = 3.16$ ← smaller

→ A2 ∈ cluster3

A3:

$d(A3, \text{seed1})= \sqrt{36} = 6$

$d(A3, \text{seed2})= \sqrt{25} = 5$ ← smaller

$d(A3, \text{seed3})= \sqrt{53} = 7.28$

→ A3 ∈ cluster2

A4:

$d(A4, \text{seed1})= \sqrt{13}$

$d(A4, \text{seed2})=0$ as A4 is seed2

$d(A4, \text{seed3})= \sqrt{52} > 0$

→ A4 ∈ cluster2

A5:

$d(A5, \text{seed1})= \sqrt{50} = 7.07$

A6:

$d(A6, \text{seed1})= \sqrt{52} = 7.21$

$$d(A5, \text{seed2}) = \sqrt{13} = 3.60 \leftarrow \text{smaller}$$

$$d(A5, \text{seed3}) = \sqrt{45} = 6.70$$

→ A5 ∈ cluster2

A7:

$$d(A7, \text{seed1}) = \sqrt{65} > 0$$

$$d(A7, \text{seed2}) = \sqrt{52} > 0$$

$$d(A7, \text{seed3}) = 0 \text{ as } A7 \text{ is seed3}$$

→ A7 ∈ cluster3

end of epoch1

$$d(A6, \text{seed2}) = \sqrt{17} = 4.12 \leftarrow \text{smaller}$$

$$d(A6, \text{seed3}) = \sqrt{29} = 5.38$$

→ A6 ∈ cluster2

A8:

$$d(A8, \text{seed1}) = \sqrt{5}$$

$$d(A8, \text{seed2}) = \sqrt{2} \leftarrow \text{smaller}$$

$$d(A8, \text{seed3}) = \sqrt{58}$$

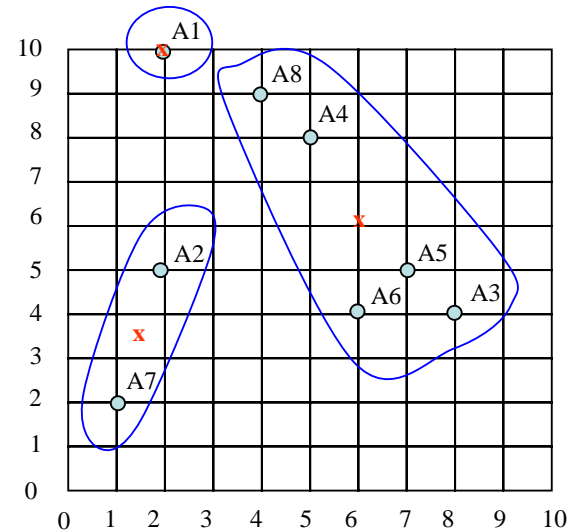
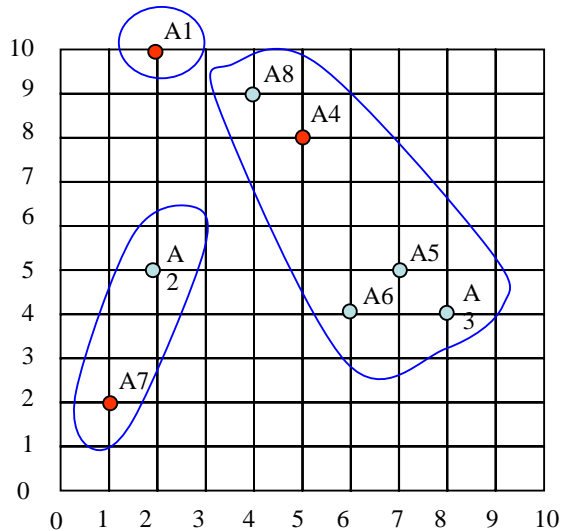
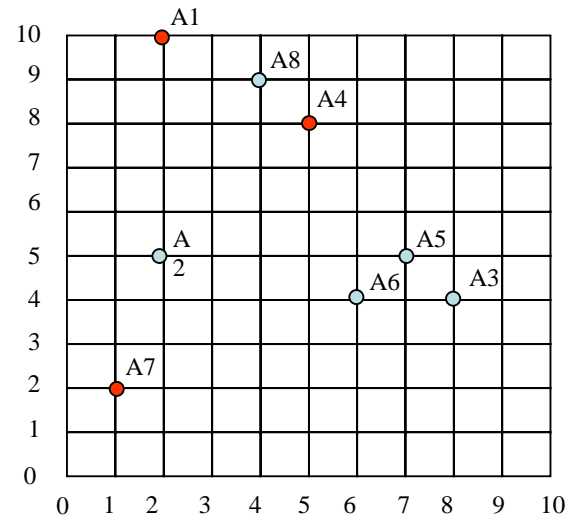
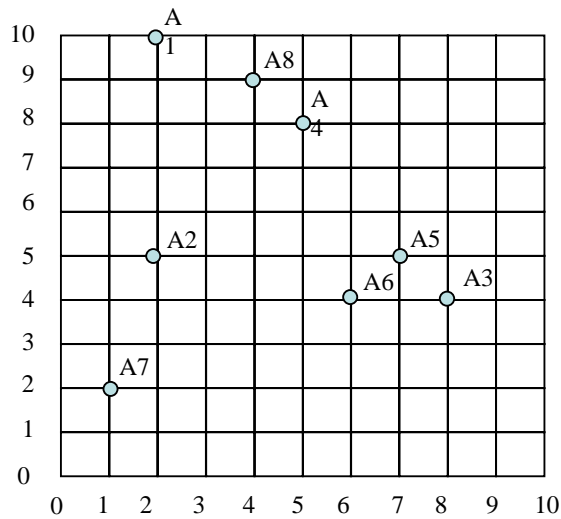
→ A8 ∈ cluster2

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

b) centers of the new clusters:

$$C1 = (2, 10), C2 = ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$$

c)



d)

We would need two more epochs. After the 2nd epoch the results would be:

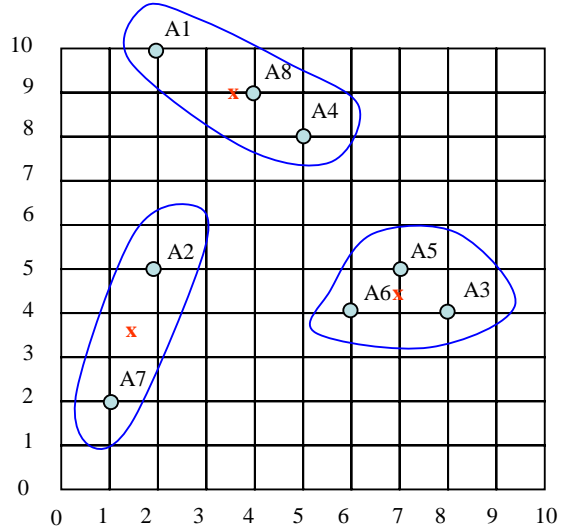
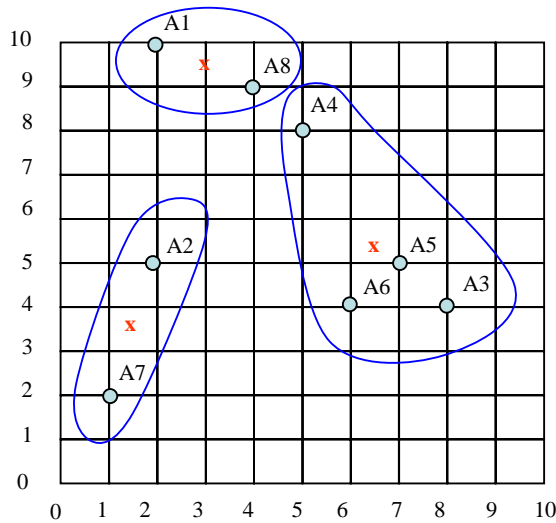
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}

with centers $C1=(3, 9.5)$, $C2=(6.5, 5.25)$ and $C3=(1.5, 3.5)$.

After the 3rd epoch, the results would be:

1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}

with centers $C1=(3.66, 9)$, $C2=(7, 4.33)$ and $C3=(1.5, 3.5)$.



Exercise 2. Nearest Neighbor clustering

Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster the examples from the previous exercise: $A1=(2,10)$, $A2=(2,5)$, $A3=(8,4)$, $A4=(5,8)$, $A5=(7,5)$, $A6=(6,4)$, $A7=(1,2)$, $A8=(4,9)$. Suppose that the threshold t is 4.

Solution:

A1 is placed in a cluster by itself, so we have $K1=\{A1\}$.

We then look at A2 if it should be added to K1 or be placed in a new cluster.

$d(A1,A2)=\sqrt{25}=5 > t \rightarrow K2=\{A2\}$

A3: we compare the distances from A3 to A1 and A2.

A3 is closer to A2 and $d(A3,A2)=\sqrt{36} > t \rightarrow K3=\{A3\}$

A4: We compare the distances from A4 to A1, A2 and A3.

A1 is the closest object and $d(A4,A1)=\sqrt{13} < t \rightarrow K1=\{A1, A4\}$

A5: We compare the distances from A5 to A1, A2, A3 and A4.

A3 is the closest object and $d(A5,A3)=\sqrt{2} < t \rightarrow K3=\{A3, A5\}$

A6: We compare the distances from A6 to A1, A2, A3, A4 and A5.

A3 is the closest object and $d(A6,A3)=\sqrt{2} < t \rightarrow K3=\{A3, A5, A6\}$

A7: We compare the distances from A7 to A1, A2, A3, A4, A5, and A6.

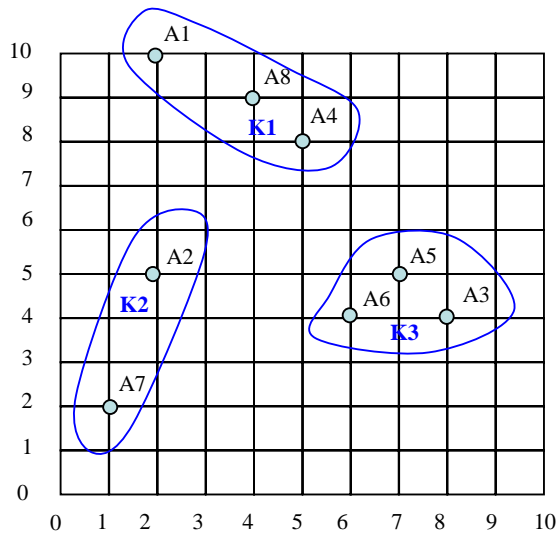
A2 is the closest object and $d(A7,A2)=\sqrt{10} < t \rightarrow K2=\{A2, A7\}$

A8: We compare the distances from A8 to A1, A2, A3, A4, A5, A6 and A7.

A4 is the closest object and $d(A8, A4) = \sqrt{2} < t \rightarrow K1 = \{A1, A4, A8\}$

Thus: $K1 = \{A1, A4, A8\}$, $K2 = \{A2, A7\}$, $K3 = \{A3, A5, A6\}$

Yes, it is the same result as with K-means.



Exercise 3. Hierarchical clustering

Use single and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms.

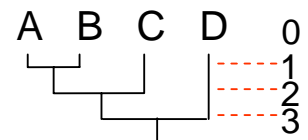
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Solution:

Agglomerative \rightarrow initially every point is a cluster of its own and we merge cluster until we end-up with one unique cluster containing all points.

a) single link: distance between two clusters is the shortest distance between a pair of elements from the two clusters.

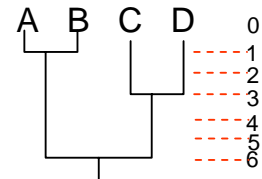
d	k	K	Comments
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	Merge {A} and {B} since A & B are the closest: $d(A, B) = 1$
2	2	{A, B, C}, {D}	Merge {A, B} and {C} since B & C are the closest: $d(B, C) = 2$
3	1	{A, B, C, D}	Merge D



b) complete link: distance between two clusters is the longest distance between a pair of elements from

the two clusters.

d	k	K	Comments
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	$d(A,B)=1 \leq 1 \rightarrow$ merge {A} and {B}
2	3	{A, B}, {C}, {D}	$d(A,C)=4 > 2$ so we can't merge C with {A,B} $d(A,D)=5 > 2$ and $d(B,D)=6 > 2$ so we can't merge D with {A, B} $d(C,D)=3 > 2$ so we can't merge C and D
3	2	{A, B}, {C, D}	- $d(A,C)=4 > 3$ so we can't merge C with {A,B} - $d(A,D)=5 > 3$ and $d(B,D)=6 > 3$ so we can't merge D with {A, B} - $d(C,D)=3 \leq 3$ so merge C and D
4	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as $d(A,D)=5 > 4$ (and also $d(B,D)=6 > 4$) although $d(A,C)=4 \leq 4$, $d(B,C)=2 \leq 4$
5	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as $d(B,D)=6 > 5$
6	1	{A, B, C, D}	{C, D} can be merged with {A, B} since $d(B,D)=6 \leq 6$, $d(A,D)=5 \leq 6$, $d(A,C)=4 \leq 6$, $d(B,C)=2 \leq 6$



Exercise 4: Hierarchical clustering (to be done at your own time, not in class)

Use single-link, complete-link, average-link agglomerative clustering as well as medoid and centroid to cluster the following 8 examples:

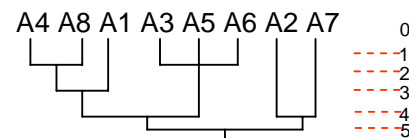
$A_1=(2,10)$, $A_2=(2,5)$, $A_3=(8,4)$, $A_4=(5,8)$, $A_5=(7,5)$, $A_6=(6,4)$, $A_7=(1,2)$, $A_8=(4,9)$.

The distance matrix is the same as the one in Exercise 1. Show the dendrograms.

Solution:

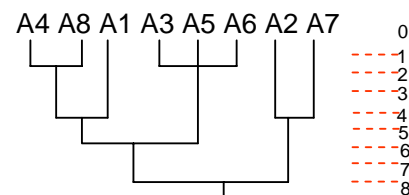
Single Link:

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}



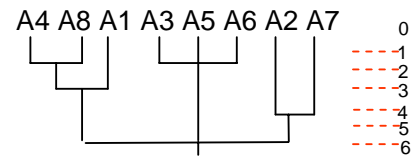
Complete Link

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
7	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
8	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average Link

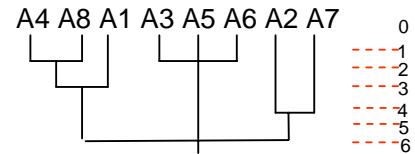
d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average distance from {A3, A5, A6} to {A1, A4, A8} is 5.53 and is 5.75 to {A2, A7}

Centroid

D	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Centroid of {A4, A8} is B=(4.5, 8.5) and centroid of {A3, A5, A6} is C=(7, 4.33)

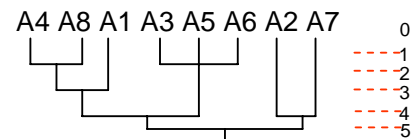
distance(A1, B) = 2.91 Centroid of {A1, A4, A8} is D=(3.66, 9) and of {A2, A7} is E=(1.5, 3.5)

distance(D,C) = 5.74 distance(D,E) = 5.90

Medoid

This is not deterministic. It can be different depending upon which medoid in a cluster we chose.

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}



Exercise 5: DBScan

If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

The distance matrix is the same as the one in Exercise 1. Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to $\sqrt{10}$?

Solution:

What is the Epsilon neighborhood of each point?

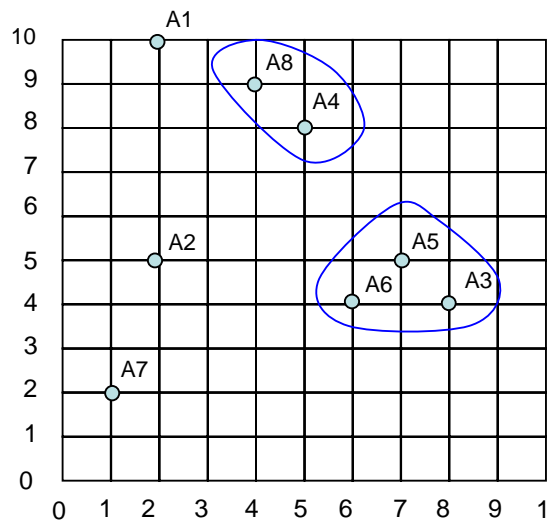
$N_2(A1)=\{\}$; $N_2(A2)=\{\}$; $N_2(A3)=\{A5, A6\}$; $N_2(A4)=\{A8\}$; $N_2(A5)=\{A3, A6\}$;

$N_2(A6)=\{A3, A5\}$; $N_2(A7)=\{\}$; $N_2(A8)=\{A4\}$

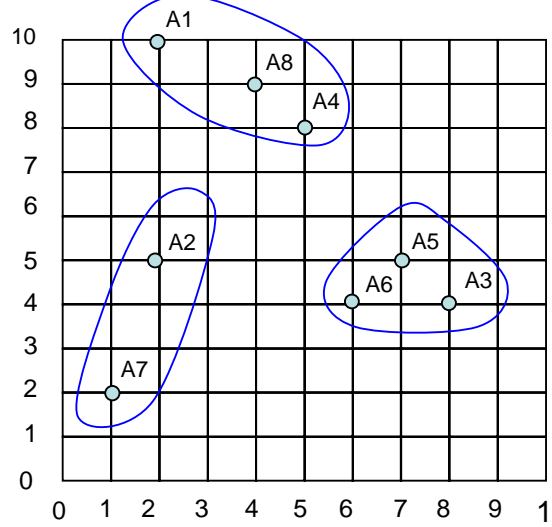
So A1, A2, and A7 are outliers, while we have two clusters $C1=\{A4, A8\}$ and $C2=\{A3, A5, A6\}$

If Epsilon is $\sqrt{10}$ then the neighborhood of some points will increase:

A1 would join the cluster C1 and A2 would joint with A7 to form cluster $C3=\{A2, A7\}$.



Epsilon = 2



Epsilon = $\sqrt{10}$