Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 9 (Unsupervised Learning - part 2)

Q1. 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

a) The new clusters (i.e. the examples belonging to each cluster)
b) The centers of the new clusters
c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
d) How many more iterations are needed to converge? Draw the result for each epoch.

Ans:

d(a,b) denotes the Eucledian 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:

BACS3013 Data Science

A1:
d(A1, seed1)=0 as A1 is seed1
d(A1, seed2)= $\sqrt{13}$ >0
d(A1, seed3)= $\sqrt{65}$ >0
➔A1 ∈ cluster1

A2:
d(A2,seed1)= $\sqrt{25}$ = 5
d(A2, seed2)= $\sqrt{18}$ = 4.24
d(A2, seed3)= $\sqrt{10}$ = 3.16 ← smaller
➔ A2 ∈ cluster3

A3:
d(A3, seed1)= $\sqrt{36}$ = 6
d(A3, seed2)= $\sqrt{25}$ = 5 ← smaller
d(A3, seed3)= $\sqrt{53}$ = 7.28
➔ A3 ∈ cluster2

A4:
d(A4, seed1)= $\sqrt{13}$
d(A4, seed2)=0 as A4 is seed2
d(A4, seed3)= $\sqrt{52}$ >0
➔ A4 ∈ cluster2

A5:
d(A5, seed1)= $\sqrt{50}$ = 7.07
d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller
d(A5, seed3)= $\sqrt{45}$ = 6.70
➔ A5 ∈ cluster2

A6:
d(A6, seed1)= $\sqrt{52}$ = 7.21
d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller
d(A6, seed3)= $\sqrt{29}$ = 5.38
➔ A6 ∈ cluster2

A7:
d(A7, seed1)= $\sqrt{65}$ >0
d(A7, seed2)= $\sqrt{52}$ >0
d(A7, seed3)=0 as A7 is seed3
➔ A7 ∈ cluster3
end of epoch1

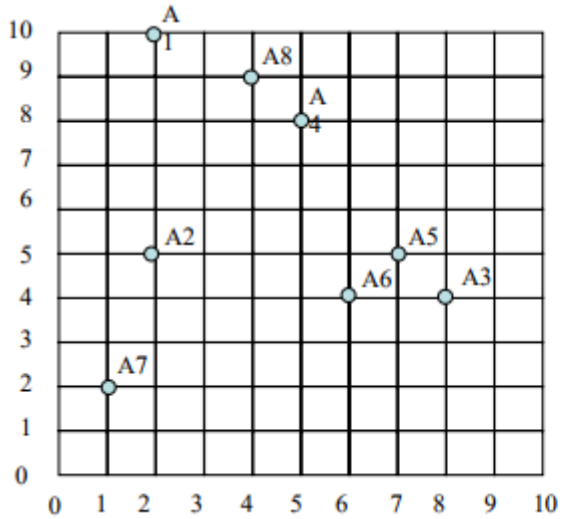
A8:
d(A8, seed1)= $\sqrt{5}$
d(A8, seed2)= $\sqrt{2}$ ← smaller
d(A8, seed3)= $\sqrt{58}$
➔ A8 ∈ cluster2

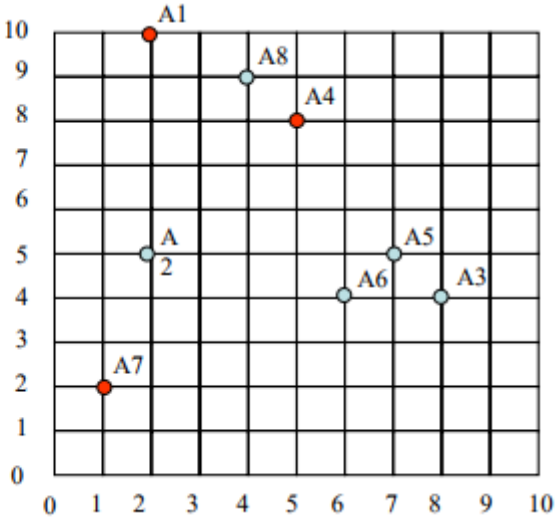
new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

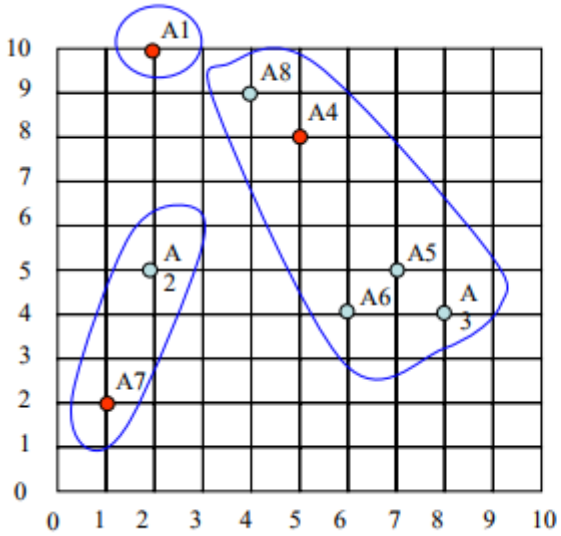
b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

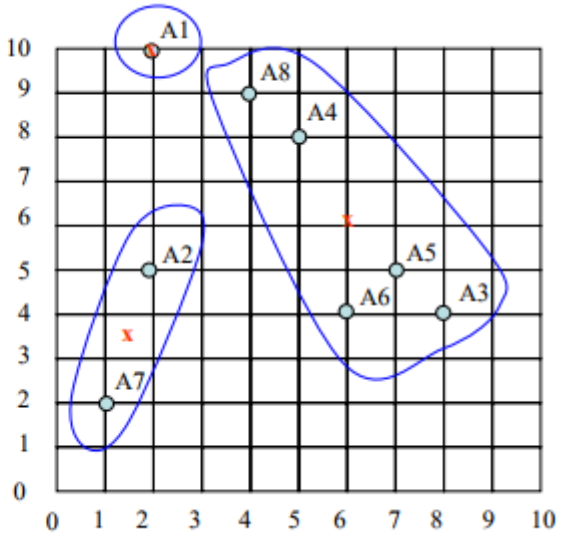
BACS3013 Data Science

c)

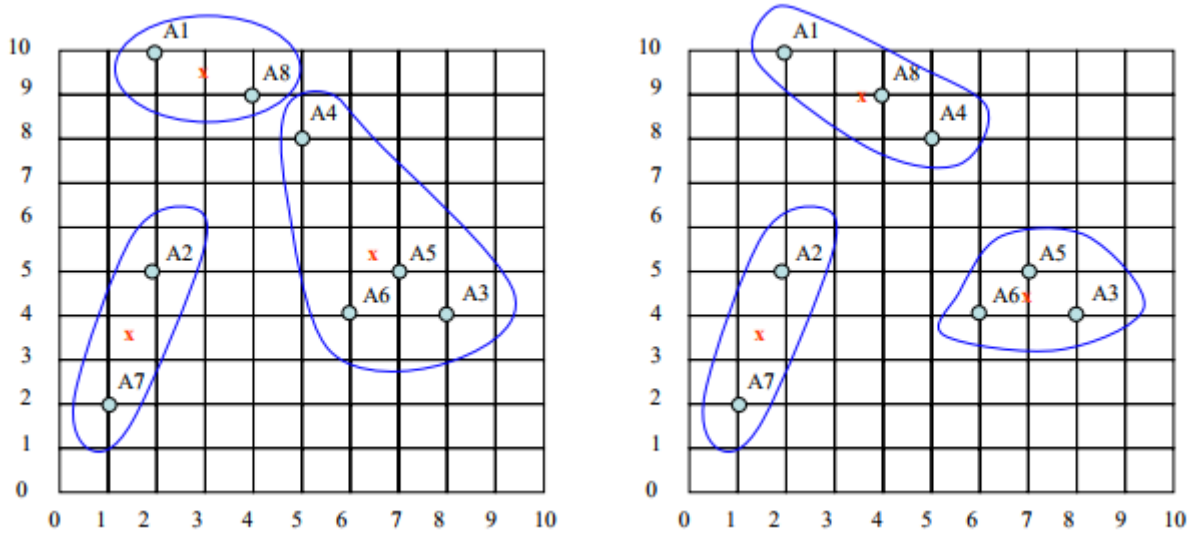








d) We would need two more epochs. After the 2nd epoch the results would be: 1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7} with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5). After the 3rd epoch, the results would be: 1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7} with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



Q2. (This question is related to DBScan)

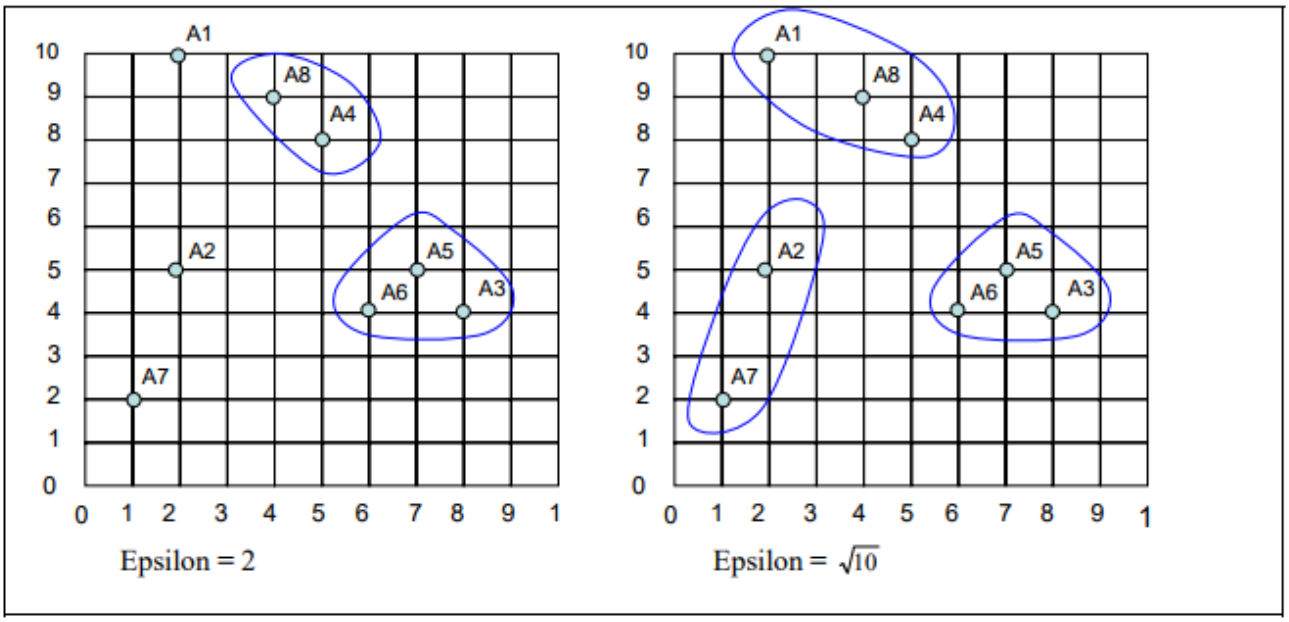
If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples:
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

The distance matrix is the same as the one in Q1. Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to 10 ?

a) k-means clustering is a method of vector quantization
b) k-means clustering aims to partition n observations into k clusters
c) k-nearest neighbor is same as k-means
d) None of the Mentioned

Ans:

What is the Epsilon neighborhood of each point?
N2(A1)={}; N2(A2)={}; N2(A3)={A5, A6}; N2(A4)={A8}; N2(A5)={A3, A6};
N2(A6)={A3, A5}; N2(A7)={}; N2(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 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}.



Q3. What are the advantages and disadvantages of DBSCAN?

Ans:

Advantages:

- **Clusters can have arbitrary shape and size**
- **Number of clusters is determined automatically**
- **Can separate clusters from surrounding noise**
- **Can be supported by spatial index structures**

Disadvantages:

- **Input parameter may be difficult to determine**
- **Very sensitive to input parameter setting**