CS481 Data Science

Assignment 4

May 11, 2020

Bilal Khan 16K3778

1 K-Means Clustering Using Euclidean Distance

Using kmeans algorithm and Euclidean distance to cluster the following 8 points into 3 clusters. Using A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2) , A8 = (4,9). Consider initial seeds as A1, A4, and A7. Run algorithm for 1 iteration only. At the end of iteration 1, show:

- The new clusters (i.e. the examples belong to each cluster)
- The centre of the new clusters
- Draw 10 x 10 space and all 8 points and show the clusters after 1st iteration and the new centroids
- Without running algorithm again, guess how many more iterations are required to converge. Draw the result of each iteration.

CS 481 Data Science Date: 10/05/2020

Assignment 4.

Q1.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	5	8.48	3.60	7.07	7.21	8.06	2.23
A_2		0	6.50	4.24	5.0	4.12	3.16	4.47
A_3			0	5	1.41	2.0	7.28	6.40
A_4				0	3.60	4.12	7.12	1.41
A_5					0	1.41	6.70	5.0
A_6						0	5.38	5.38
A_7							0	7.61
A_8								0

Distance has been calculated using Euclidean distance.

$$\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Considering distance for seeds A_1 A_4 and A_7.
New clusters would be.

(A_1) A_3 (A_4) A_5 A_6 A_8 A_2 (A_7)

Center of new clusters

1.) $(A_1) = (2, 10)$ 2.) $\left(\frac{\sum_{i=0}^{n} X_i}{n}, \frac{\sum_{i=0}^{n} Y_i}{n}\right)$

$= (6, 6)$

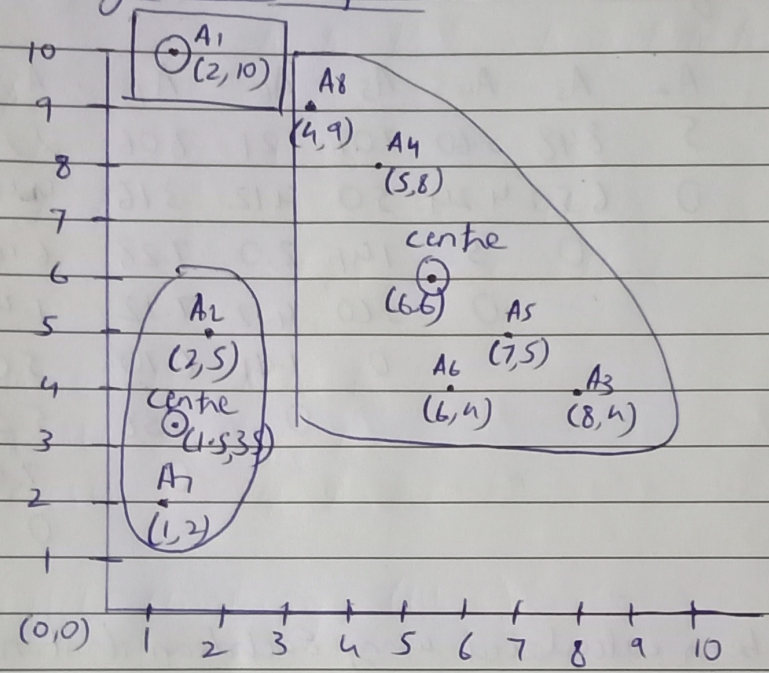
3.) $(1.5, 3.5)$

(1)

Figure 1: Clustering iteration and centroid

Date: 10/05/202

Drawing in a 2D space:



When will it all converge?

Given from the prespective of the centres and far they are plus from having an idea from the question to solve ahead of this assignment I'd say they'd converge in about 8 iterations

Figure 2: Drawn in 10x10 space

2 Hierarchical Clustering

Using hierarchical clustering algorithms (Single, Complete, Group Average and Distance b/w centroids) and Euclidean distance to cluster the following 8 points into 3 clusters. Using A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2) , A8 = (4,9).

Date: 10/05/2020

Q2. Single hierarchical clustering (Min)

Continuing from the distance table in question 1.

2nd iteration

	A_1	A_2	A_3A_5	A_4	A_6	A_7	A_8
A_1	0						
A_2	5	0					
A_3A_5	7.07	5	0				
A_4	3.60	4.24	3.60	0			
A_6	7.21	4.12	(1.41)	4.12	0		
A_7	8.06	3.13	6.70	7.21	5.37	0	
A_8	2.23	4.47	5	(1.41)	5.37	7.61	0

3rd iteration

	A_1	A_2	$A_3A_5A_6$	A_4A_8	A_7
A_1	0				
A_2	5	0			
$A_3A_5A_6$	7.071	4.12	0		
A_4A_8	2.236	4.24	3.60	0	
A_7	8.06	3.16	5.38	7.21	0

URBANE PAPER PRODUCT

(3)

Figure 3: Single linkage

Date: 10/05/2020

Fourth Iteration.

	A_1, A_4, A_8	A_2	A_3, A_5, A_6	A_7
$A_1 A_4 A_8$	0			
A_2	4.25	0		
$A_3 A_5 A_6$	3.60	4.12	0	
A_7	7.21	3.16	5.38	0

Cluster 1 = $A_1 A_4 A_8$
Cluster 2 = $A_2 A_7$
Cluster 3 = $A_3 A_5 A_6$

Complete Clustering (Max)

1st iteration.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0							
A_2	5	0						
A_3	8.49	6.08	0					
A_4	3.60	4.24	5	0				
A_5	7.07	5	1.41	3.60	0			
A_6	7.21	4.12	2	4.12	1.41	0		
A_7	8.06	3.16	7.28	7.21	6.7	5.38	0	
A_8	2.24	4.47	6.4	1.41	5	5.38	7.60	0

Figure 4: Complete linkage

Date:

2nd iteration.

	A_1	A_2	A_3 A_5	A_4	A_6	A_7	A_8
A_1	0						
A_2	5	0					
A_3A_5	8.49	6.08	0				
A_4	3.6	4.24	5	0			
A_6	7.21	4.12	2	4.12	0		
A_7	8.06	3.16	7.28	7.21	5.38	0	
A_8	2.24	4.47	6.4	(4.47)	5.38	7.60	0

3rd iteration.

	A_1	A_2	A_3A_5	A_4A_8	A_6	A_7
A_1	0					
A_2	5	0				
A_3A_5	8.48	6.08	0			
A_4A_8	3.6	4.47	6.4	0		
A_6	7.21	4.12	(2)	5.38	0	
A_7	8.06	3.16	7.28	7.6	5.38	0

4TH iteration

	A_1	A_2	A_3 A_5 A_6	A_4A_8	A_7
A_1	0				
A_2	5	0			
$A_3A_5A_6$	8.49	3.16	0		
A_4A_8	3.6	4.47	6.4	0	
A_7	8.06	(3.16)	7.28	7.6	0

(5)

Figure 5: Complete linkage

Date: 10/05/20

5TH iteration

	A_1	A_2A_7	$A_3A_5A_6$	A_4A_8
A_1	0			
A_2A_7	8.06	0		
$A_3A_5A_6$	8.49	7.28	0	
A_4A_8	3.60	7.60	6.4.	0

Cluster1 = $A_1A_4A_8$
Cluster2 = A_2A_7
Cluster3 = $A_3A_5A_6$

Average Linkage

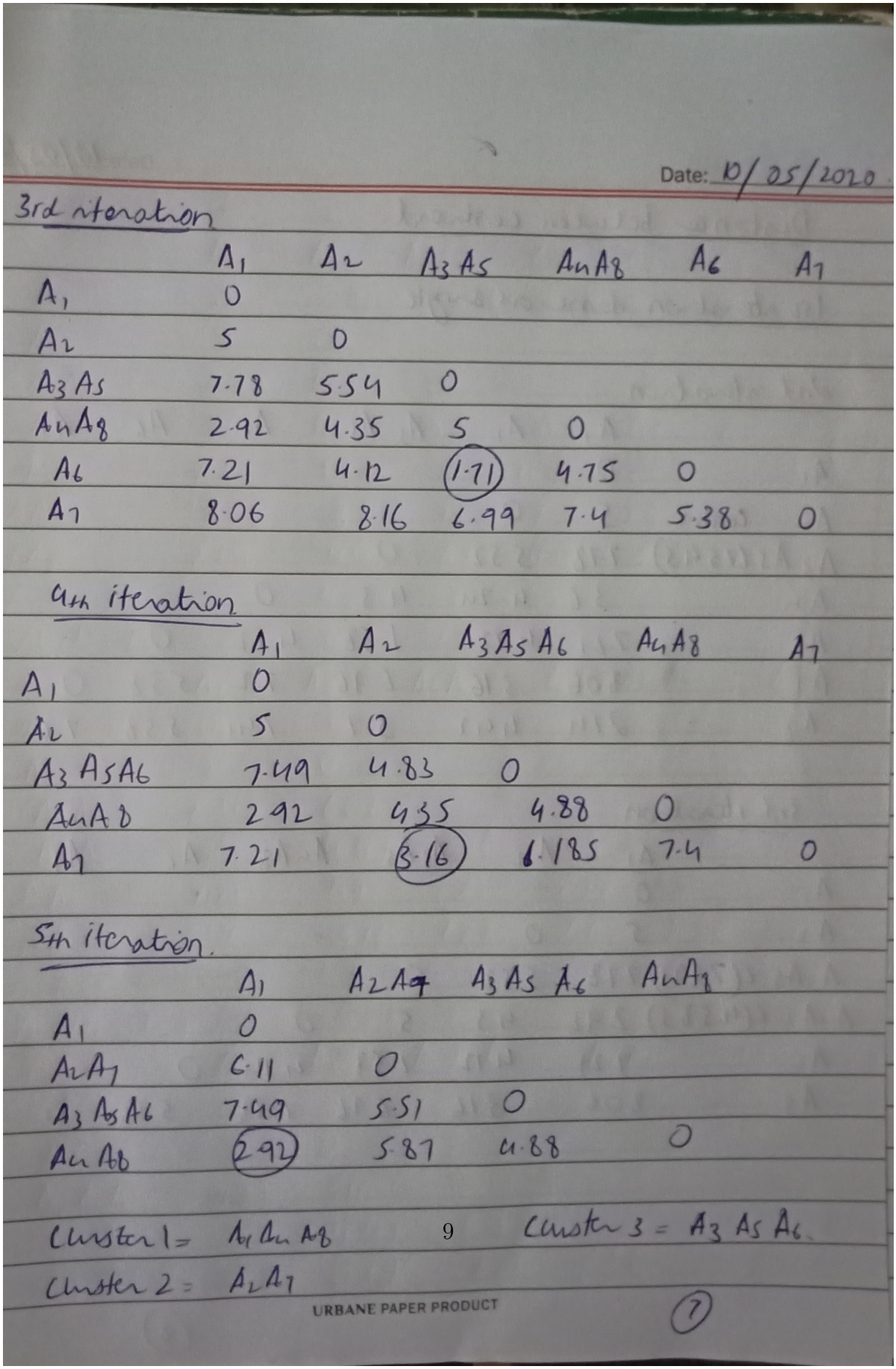
Step 1 same all before.

2nd iteration

	A_1	A_2	A_3A_5	A_4	A_6	A_7	A_8
A_1	0						
A_2	5	0					
A_3A_5	7.78	5.54	0				
A_4	3.6	4.24	5.43	0			
A_6	7.21	4.12	1.71	4.12	0		
A_7	8.06	3.16	6.99	7.2	5.38	0	
A_8	2.24	4.47	5.9	1.41	5.38	7.6	0

(6)

Figure 6: Average linkage


Date: 10/05/2020

3rd iteration

	A_1	A_2	A_3A_5	A_4A_8	A_6	A_7
A_1	0					
A_2	5	0				
A_3A_5	7.78	5.54	0			
A_4A_8	2.92	4.35	5	0		
A_6	7.21	4.12	(1.71)	4.75	0	
A_7	8.06	8.16	6.99	7.4	5.38	0

4th iteration

	A_1	A_2	$A_3A_5A_6$	A_4A_8	A_7
A_1	0				
A_2	5	0			
$A_3A_5A_6$	7.49	4.83	0		
A_4A_8	2.92	4.35	4.88	0	
A_7	7.21	(3.16)	6.185	7.4	0

5th iteration.

	A_1	A_2A_7	$A_3A_5A_6$	A_4A_8
A_1	0			
A_2A_7	6.11	0		
$A_3A_5A_6$	7.49	5.51	0	
A_4A_8	(2.92)	5.87	4.88	0

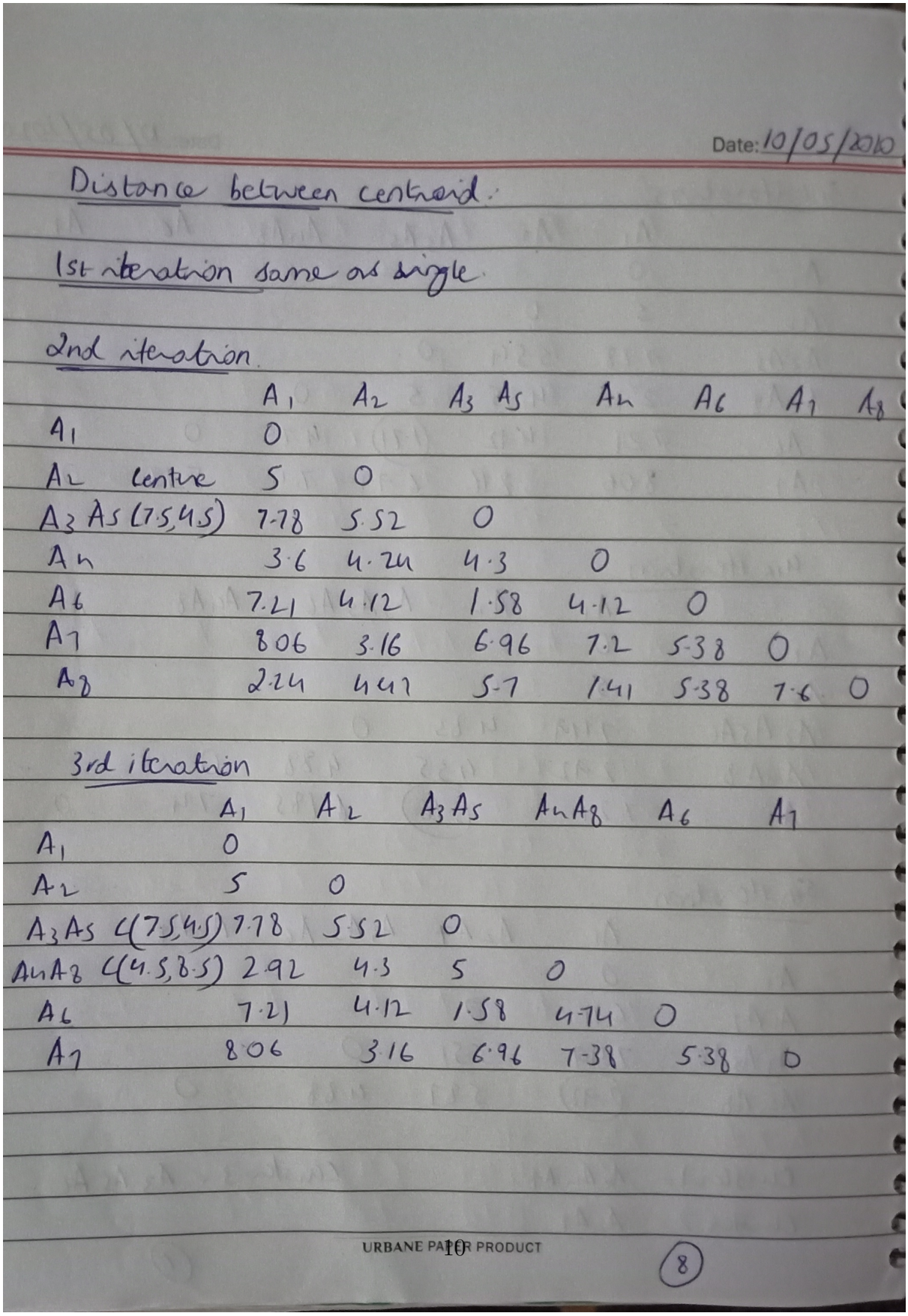
Cluster 1 = $A_1A_4A_8$ 9 Cluster 3 = $A_3A_5A_6$

Cluster 2 = A_2A_7

URBANE PAPER PRODUCT

(7)

Figure 7: Average linkage



Date: 10/05/2010

Distance between centroid:

1st iteration same as single

2nd iteration

	A_1	A_2	A_3 A_5	A_4	A_6	A_7	A_8
A_1	0						
A_2 centre	5	0					
A_3 A_5 (7.5,4.5)	7.78	5.52	0				
A_4	3.6	4.24	4.3	0			
A_6	7.21	4.12	1.58	4.12	0		
A_7	8.06	3.16	6.96	7.2	5.38	0	
A_8	2.24	4.47	5.7	1.41	5.38	7.6	0

3rd iteration

	A_1	A_2	A_3 A_5	A_4 A_8	A_6	A_7
A_1	0					
A_2	5	0				
A_3 A_5 (7.5,4.5)	7.78	5.52	0			
A_4 A_8 (4.5,3.5)	2.92	4.3	5	0		
A_6	7.21	4.12	1.58	4.74	0	
A_7	8.06	3.16	6.96	7.38	5.38	0

URBANE PAPER PRODUCT

(8)

Figure 8: Distance between centroid

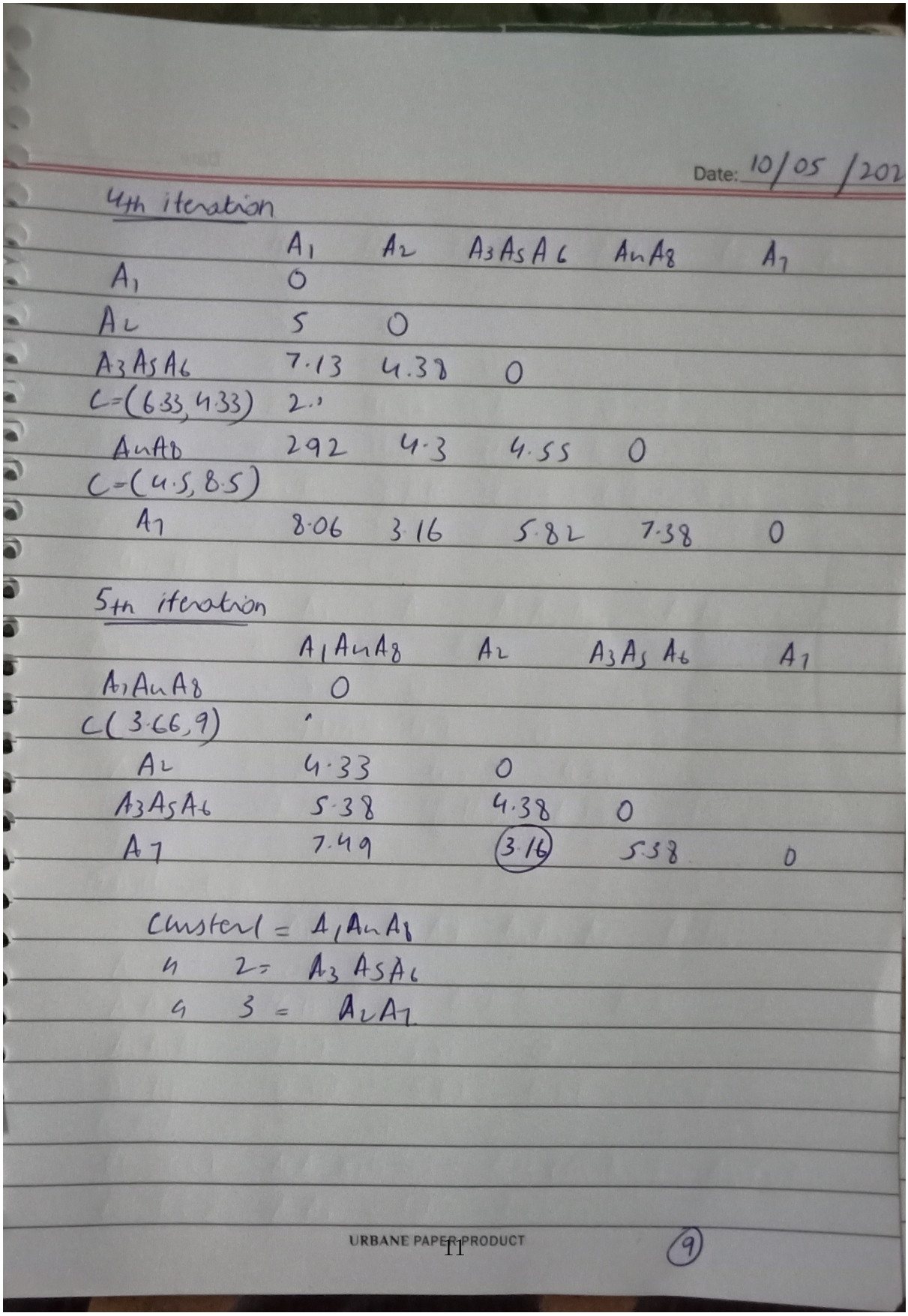


Figure 9: Distance between centroid

3 Paper Summary

Review a paper of your choice on Soft Clustering e.g. Fuzzy Kmeans (One Page Summary).

The following is a summary from the paper *Fuzzy C- Means Algorithm- A review* by *R.Suganya and R.Shanthi* it can be cited from the following link, `http://www.ijsrp.org/research-paper-1112/ijsrp-p1168.pdf`. The paper describes in an informative way how there are 2 kinds of clustering, **Hard Clustering** the one which is usually opted for where each and every points is clustered in to a single group. On the other hand is **Soft Clustering** where every single point can be selected in to multiple groups or clusters depending on the criteria and means of clustering or feature selection. *Fuzzy C Means* or FCM for short is a part of classical fuzzy clustering algorithms and this paper discusses it in detail and explains all the mathematical happenings of this algorithm. It is the most commonly used hard clustering algorithms present it was developed by *Joe Dunn* in 1974 for a special case of $K = 2$. Instead of classifying a point to a single special cluster, the objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. Some Pros and cons by the paper are:

1. Advantages
   (a) Unsupervised
   (b) Converges
2. Limitations
   (a) Long computational time
   (b) Sensitivity to the initial guess (speed,local minima)
   (c) Sensitivity to noise (outliers, NaN values)

To overcome the difficulties presented by FCM another variant of the algorithm *Probabilistic C Means* was presented which by the name proves that it works on the probabilities of distant points and keeps calculating clusters until each and every point has at least one cluster to represent with the probability of them being the part of that cluster a minimum of 0.5.It improves on the previous approach by being able to cluster noisy data more accurately and efficiently whereas the disadvantage of being sensitive to initialization is still present. This model was further perfected by means of *Fuzzy Probabilistic C Means Algorithm* aka FPCM. It utilizes both, the typicality or feature extractions of fuzzy means plus the probability model of the possibilistic model. It ignores the noisy sensitivity deficiency of FCM.