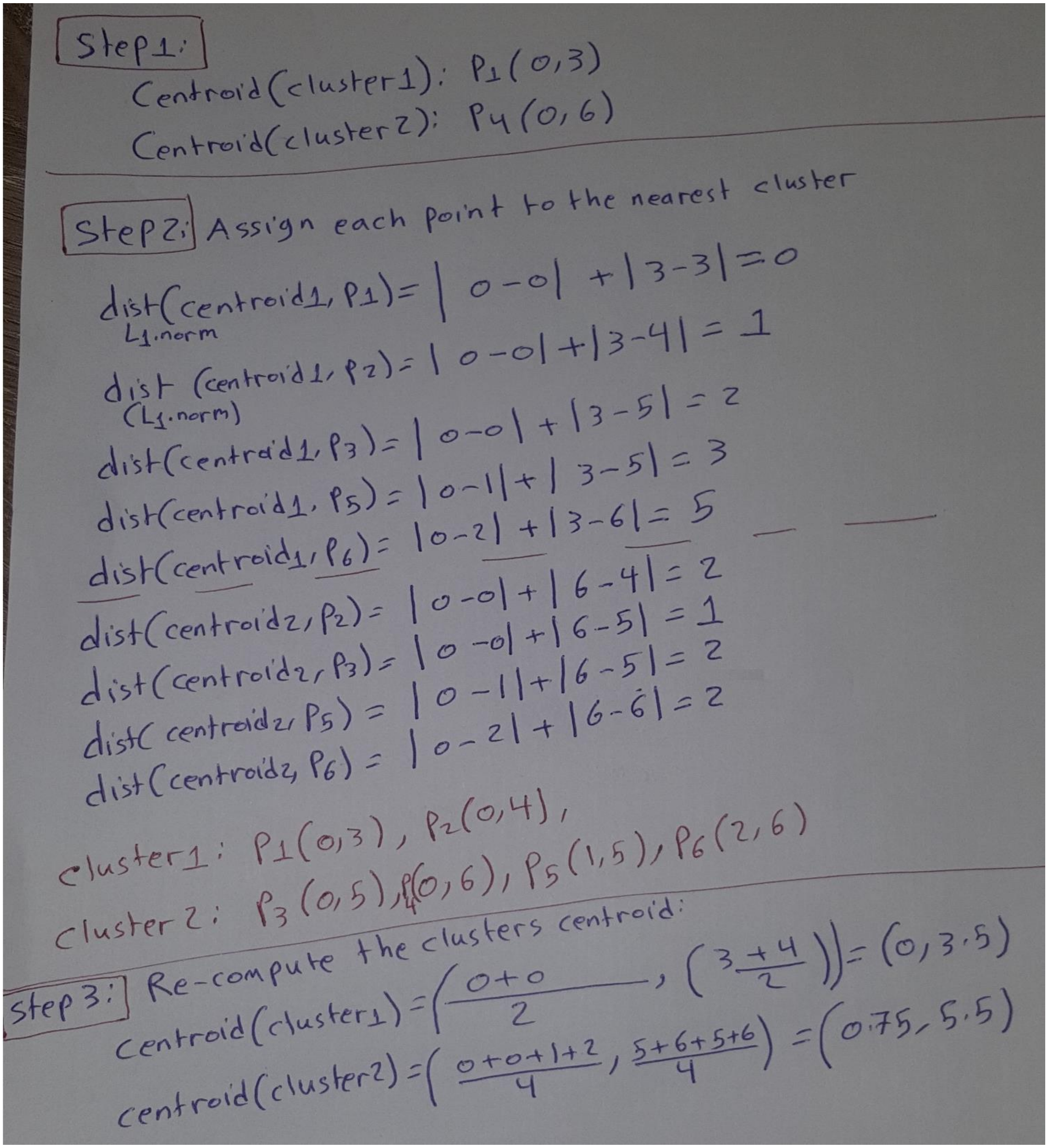
Data Clustering:

Q1) Consider the following points (p1-p6) whose coordinates are given below which belong to a two-dimensional vector space. The dataset containing these points will be clustered into 2 clusters using the ***K-means***. P1 and p4 are randomly chosen as the initial centroids. The Euclidean distance will be used as the "closeness" measure to determine the centroid each point is "close" to.

P1	P2	P3	P4	P5	P6
(0,3)	(0,4)	(0,5)	(0,6)	(1,5)	(2,6)

A) Find the cluster of each point. Show your work clearly for both cases. (15 points)



Step1:

Centroid(cluster1): $P_1(0,3)$

Centroid(cluster2): $P_4(0,6)$

Step2: Assign each point to the nearest cluster

$\underset{L_1.norm}{dist}(centroid1, P_1) = |0-0| + |3-3| = 0$

$\underset{(L_1.norm)}{dist}(centroid1, P_2) = |0-0| + |3-4| = 1$

$dist(centroid1, P_3) = |0-0| + |3-5| = 2$

$dist(centroid1, P_5) = |0-1| + |3-5| = 3$

$dist(centroid1, P_6) = |0-2| + |3-6| = 5$

$dist(centroid2, P_2) = |0-0| + |6-4| = 2$

$dist(centroid2, P_3) = |0-0| + |6-5| = 1$

$dist(centroid2, P_5) = |0-1| + |6-5| = 2$

$dist(centroid2, P_6) = |0-2| + |6-6| = 2$

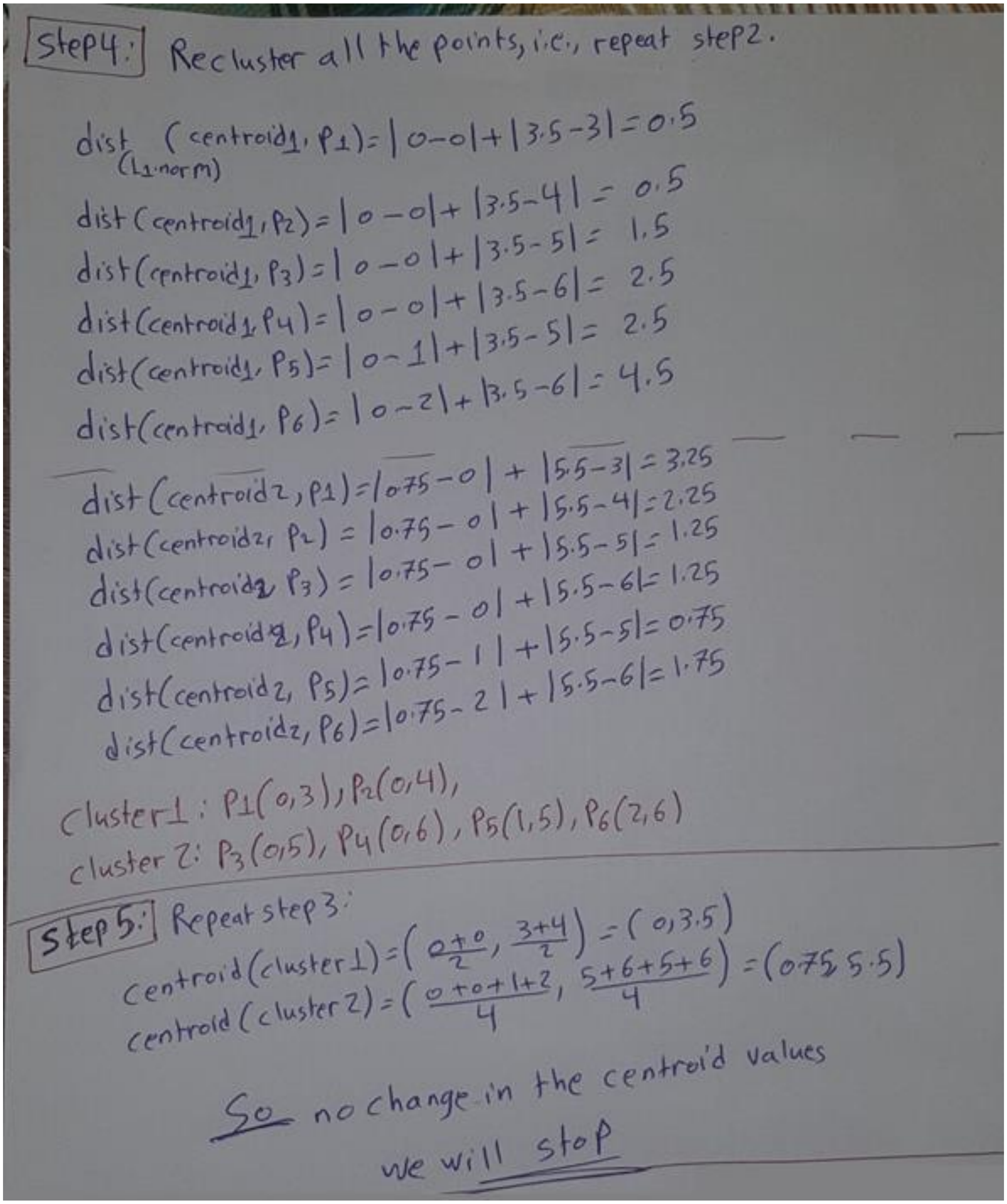
cluster1: $P_1(0,3)$, $P_2(0,4)$,

cluster 2: $P_3(0,5)$, $P_4(0,6)$, $P_5(1,5)$, $P_6(2,6)$

Step 3: Re-compute the clusters centroid:

$centroid(cluster_1) = \left(\frac{0+0}{2}, \left(\frac{3+4}{2}\right)\right) = (0, 3.5)$

$centroid(cluster2) = \left(\frac{0+0+1+2}{4}, \frac{5+6+5+6}{4}\right) = (0.75, 5.5)$



Step4: Recluster all the points, i.e., repeat step2.

$\text{dist}_{(L_1\text{norm})}(\text{centroid}_1, P_1) = |0-0| + |3.5-3| = 0.5$

$\text{dist}(\text{centroid}_1, P_2) = |0-0| + |3.5-4| = 0.5$

$\text{dist}(\text{centroid}_1, P_3) = |0-0| + |3.5-5| = 1.5$

$\text{dist}(\text{centroid}_1, P_4) = |0-0| + |3.5-6| = 2.5$

$\text{dist}(\text{centroid}_1, P_5) = |0-1| + |3.5-5| = 2.5$

$\text{dist}(\text{centroid}_1, P_6) = |0-2| + |3.5-6| = 4.5$

$\text{dist}(\text{centroid}_2, P_1) = |0.75-0| + |5.5-3| = 3.25$

$\text{dist}(\text{centroid}_2, P_2) = |0.75-0| + |5.5-4| = 2.25$

$\text{dist}(\text{centroid}_2, P_3) = |0.75-0| + |5.5-5| = 1.25$

$\text{dist}(\text{centroid}_2, P_4) = |0.75-0| + |5.5-6| = 1.25$

$\text{dist}(\text{centroid}_2, P_5) = |0.75-1| + |5.5-5| = 0.75$

$\text{dist}(\text{centroid}_2, P_6) = |0.75-2| + |5.5-6| = 1.75$

Cluster1: $P_1(0,3)$, $P_2(0,4)$,

cluster 2: $P_3(0,5)$, $P_4(0,6)$, $P_5(1,5)$, $P_6(2,6)$

Step 5: Repeat step 3:

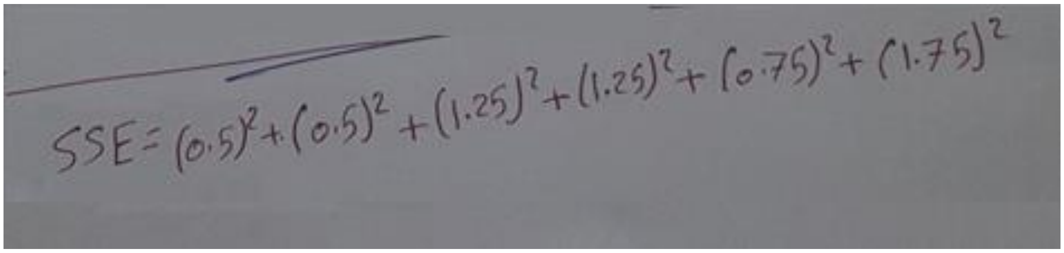
$\text{centroid}(\text{cluster}1) = \left(\frac{0+0}{2}, \frac{3+4}{2}\right) = (0, 3.5)$

$\text{centroid}(\text{cluster}2) = \left(\frac{0+0+1+2}{4}, \frac{5+6+5+6}{4}\right) = (0.75\ 5.5)$

So no change in the centroid values

we will stop

B) Calculate the Sum of Squared Error (SSE) for the two clusters found in part (A). *(7 pts)*
Note that the centroid of each cluster is found by calculating the average of all points that belong to that cluster.

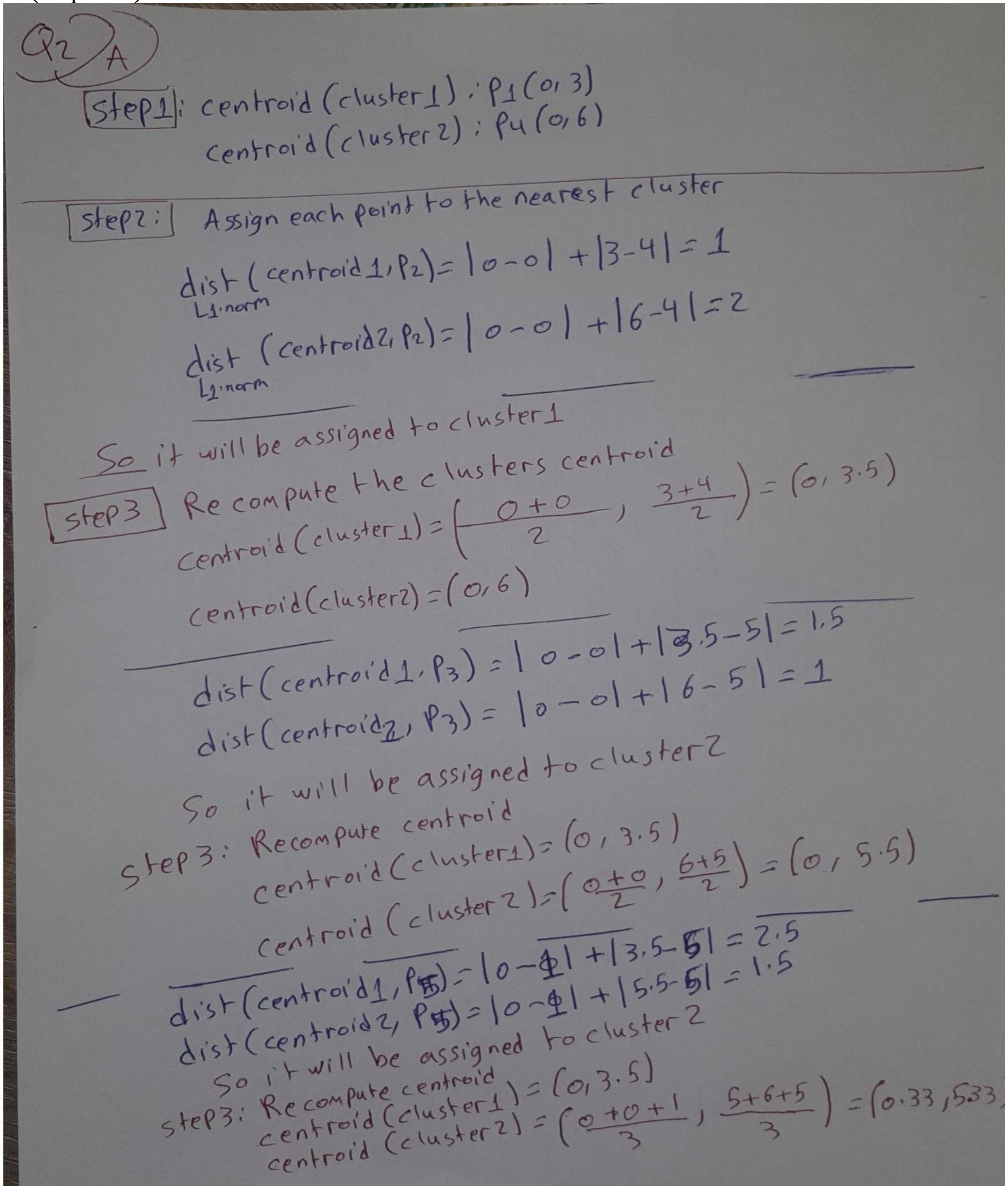


$SSE = (0.5)^2 + (0.5)^2 + (1.25)^2 + (1.25)^2 + (0.75)^2 + (1.75)^2$

<u>Q2)</u> Consider the following points (p1-p6) whose coordinates are given below which belong to a two-dimensional vector space. The dataset containing these points will be clustered into 2 clusters using the ***K-means (incremental approach)***. P1 and p4 are randomly chosen as the initial centroids. The Euclidean distance will be used as the "closeness" measure to determine the centroid each point is "close" to.

P1	P2	P3	P4	P5	P6
(0,3)	(0,4)	(0,5)	(0,6)	(1,5)	(2,6)

A) Find the cluster of each point after the first iteration of the algorithm. Show your work clearly for both cases. (15 points)



Q2 A

Step1: centroid (cluster1) : P1 (0, 3)
centroid (cluster 2) : P4 (0, 6)

Step2: Assign each point to the nearest cluster

$dist(centroid1, P_2) = |0-0| + |3-4| = 1$
L1.norm

$dist(centroid2, P_2) = |0-0| + |6-4| = 2$
L1.norm

So it will be assigned to cluster1

step3 Re compute the clusters centroid

$centroid(cluster1) = \left(\frac{0+0}{2}, \frac{3+4}{2}\right) = (0, 3.5)$

$centroid(cluster2) = (0, 6)$

$dist(centroid1, P_3) = |0-0| + |3.5-5| = 1.5$

$dist(centroid2, P_3) = |0-0| + |6-5| = 1$

So it will be assigned to cluster2

step 3: Recompute centroid

$centroid(cluster1) = (0, 3.5)$

$centroid(cluster2) = \left(\frac{0+0}{2}, \frac{6+5}{2}\right) = (0, 5.5)$

$dist(centroid1, P_5) = |0-1| + |3.5-5| = 2.5$

$dist(centroid2, P_5) = |0-1| + |5.5-5| = 1.5$

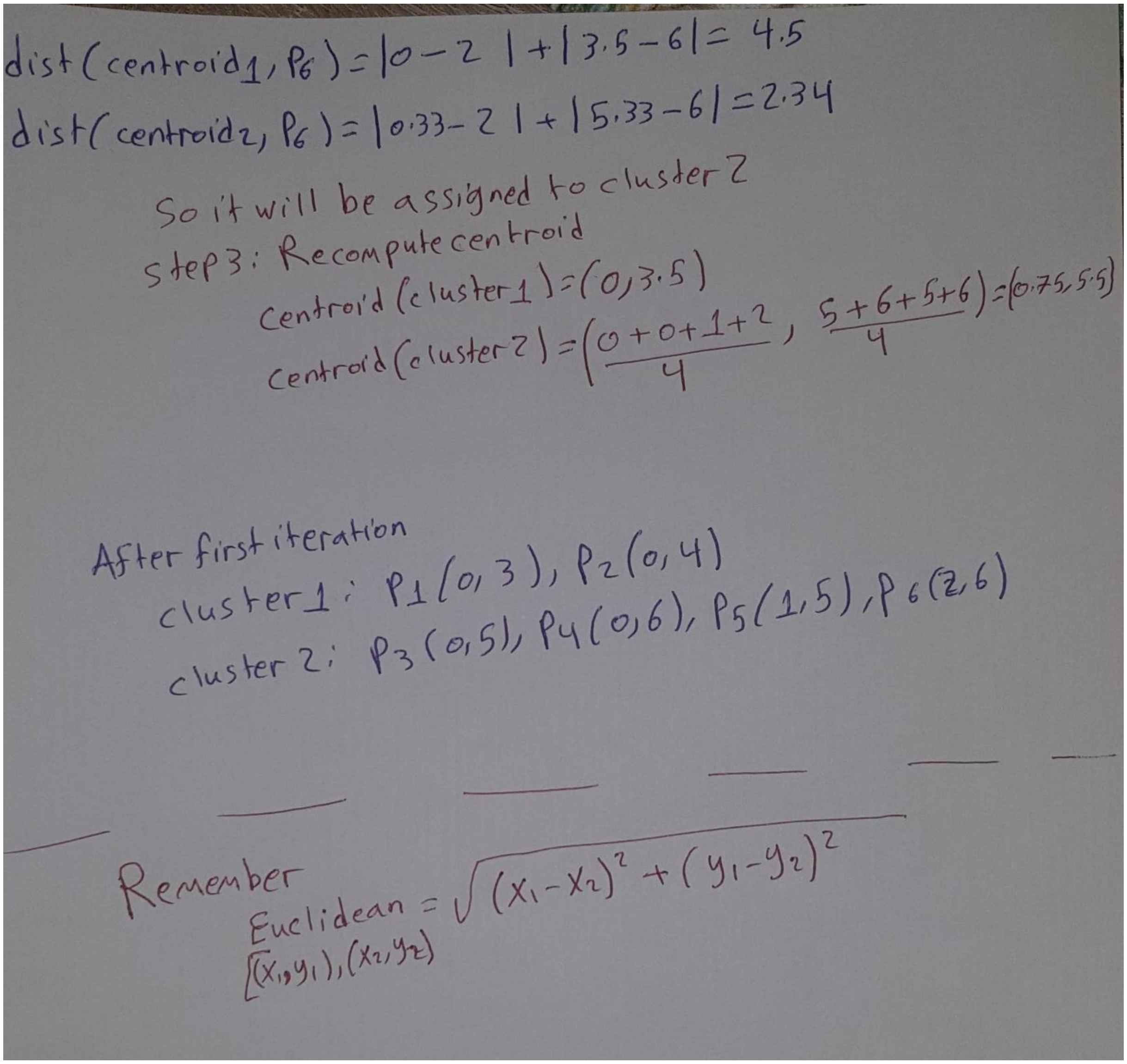
So it will be assigned to cluster 2

step3: Recompute centroid

$centroid(cluster1) = (0, 3.5)$

$centroid(cluster2) = \left(\frac{0+0+1}{3}, \frac{5+6+5}{3}\right) = (0.33, 5.33$

centroid



$\text{dist}(\text{centroid}_1, P_6) = |0 - 2| + |3.5 - 6| = 4.5$

$\text{dist}(\text{centroid}_2, P_6) = |0.33 - 2| + |5.33 - 6| = 2.34$

So it will be assigned to cluster 2

step 3: Recompute centroid

$\text{centroid}(\text{cluster}_1) = (0, 3.5)$

$\text{centroid}(\text{cluster}2) = \left(\frac{0+0+1+2}{4}, \frac{5+6+5+6}{4}\right) = (0.75, 5.5)$

After first iteration

cluster 1: $P_1(0,3), P_2(0,4)$

cluster 2: $P_3(0,5), P_4(0,6), P_5(1,5), P_6(2,6)$

Remember

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

$[(x_1, y_1), (x_2, y_2)$

B) Calculate the Sum of Squared Error (SSE) for the two clusters found in part (A). *(7 pts)*

Note that the centroid of each cluster is found by calculating the average of all points that belong to that cluster.