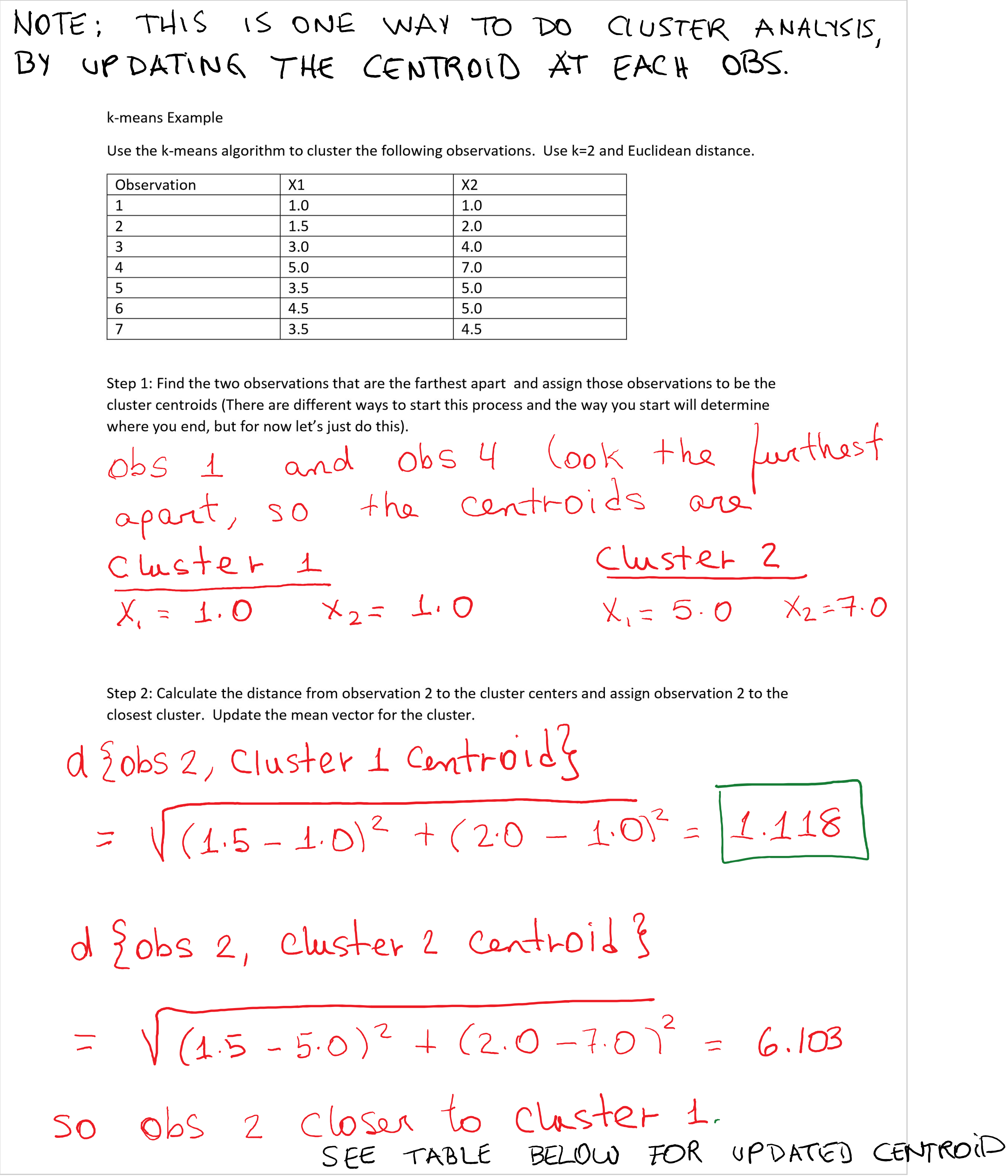
Printout

Wednesday, March 7, 2018 12:31 PM



NOTE: THIS IS ONE WAY TO DO CLUSTER ANALYSIS, BY UPDATING THE CENTROID AT EACH OBS.

k-means Example

Use the k-means algorithm to cluster the following observations. Use k=2 and Euclidean distance.

Observation	X1	X2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1: Find the two observations that are the farthest apart and assign those observations to be the cluster centroids (There are different ways to start this process and the way you start will determine where you end, but for now let's just do this).

obs 1 and obs 4 look the furthest apart, so the centroids are

Cluster 1		Cluster 2	
$X_1 = 1.0$	$X_2 = 1.0$	$X_1 = 5.0$	$X_2 = 7.0$

Step 2: Calculate the distance from observation 2 to the cluster centers and assign observation 2 to the closest cluster. Update the mean vector for the cluster.

d {obs 2, Cluster 1 Centroid}

$$= \sqrt{(1.5 - 1.0)^2 + (2.0 - 1.0)^2} = \boxed{1.118}$$

d {obs 2, cluster 2 centroid}

$$= \sqrt{(1.5 - 5.0)^2 + (2.0 - 7.0)^2} = 6.103$$

so obs 2 closer to cluster 1.

SEE TABLE BELOW FOR UPDATED CENTROID

Step 3: Calculate the distance from observation 3 to each cluster and assign it to the closest cluster. Update the mean vector for that cluster.

$d\{obs\ 3,\ Cluster\ 1\ centroid\} =$

$$\sqrt{(3-1.25)^2 + (4-1.5)^2} = \boxed{3.052}$$

Cluster 1

SEE UPDATED CENTROID

$d\{obs\ 3,\ cluster\ 2\ centroid\} =$

$$\sqrt{(3-5)^2 + (4-7)^2} = 3.606$$

Step 4: Calculate the distance from observation 5 to each cluster and assign it to the closest cluster. Update the mean vector for that cluster.

$d\{obs\ 5,\ cluster\ 1\ centroid\} =$

$$\sqrt{(3.5-1.83)^2 + (5-2.33)^2} = 3.149$$

$d\{obs\ 5,\ cluster\ 2\ centroid\} =$

$$\sqrt{(3.5-5)^2 + (5-7)^2} = \boxed{2.5}$$

Cluster 2

SEE UPDATED CENTROID BELOW

Step 5: Calculate the distance from observation 6 to each cluster and assign it to the closest cluster. Update the mean vector for that cluster.

$d\{obs\ 6,\ cluster\ 1\ centroid\} =$

$$\sqrt{(4.5-1.83)^2 + (5-2.33)^2} = 3.776$$

$d\{obs\ 6,\ cluster\ 2\ centroid\} =$

$$\sqrt{(4.5-4.25)^2 + (5-6)^2} = \boxed{1.031}$$

Cluster 2

SEE UPDATES IN TABLE

Step 6: Calculate the distance from observation 7 to each cluster and assign it to the closest cluster. Update the mean vector for that cluster.

$d\{\text{obs } 7, \text{ cluster } 1 \text{ centroid}\} =$

$$\sqrt{(3.5-1.83)^2 + (4.5-2.33)^2} = 2.738$$

$d\{\text{obs } 7, \text{ cluster } 2 \text{ centroid}\} =$

$$\sqrt{(3.5-4.33)^2 + (4.5-5.67)^2} = \boxed{1.434}$$ Cluster 2

Fill in the table with a summary of your steps 1-6 from above. I filled in step 1 for you.

	Cluster 1		Cluster 2	
Step	Obs.	Mean Vector 1	Obs.	Mean Vector 2
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	2	(1.25, 1.5)	5	(4.25, 6)
3	3	(1.83, 2.33)	6	(4.33, 5.67)
4			7	(4.13, 5.38)
5				
6				

Now we have to check and make sure that each observation is properly assigned. We will compare each observation's distance to its own cluster and the other cluster. If an observation is in the wrong cluster, re-assign it and re-compute the cluster center. Continue this process until nothing can be moved.

Observation	Distance to Mean Vector 1	Distance to Mean Vector 2
1	1.568	5.378
2	0.467	4.277
3	2.039	1.777
4	5.644	1.844
5	3.149	0.727
6	3.776	0.537
7	2.738	1.076

should be assigned to Cluster 2

Compute the final mean vectors and clusters.

Cluster 1: (1.25, 1.5)

Cluster2: (3.9, 5.1)