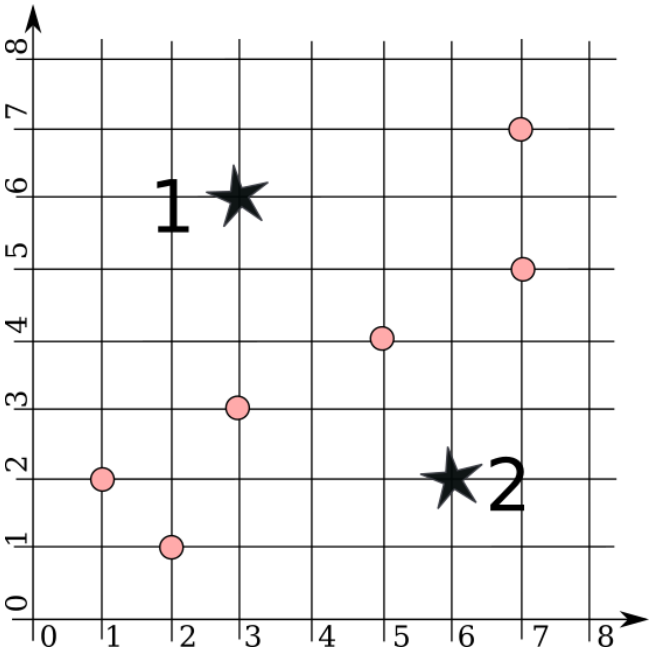
ANSWERS

Exercise 1 - k-Means Clustering ANSWER

i Determining the clusters

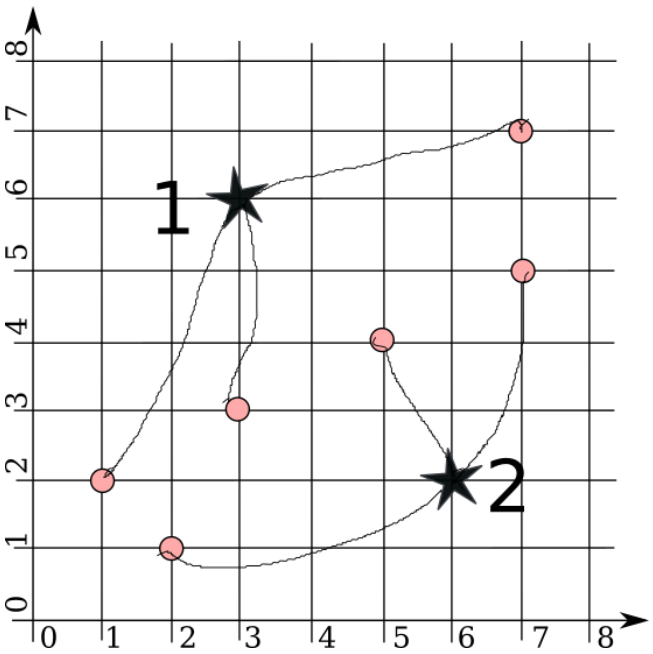
The iterative k-Means Clustering method consists of two steps that are repeated until the clusters are defined: (1) association of instances with clusters and (2) centroid calculation.

As the initial cluster-representative points have been given, we just have to label them. Let's call the cluster represented by the upper-left star **cluster 1** and the other one **cluster 2**.



*****

Now we start with step 1. We associate each point with the cluster-representative point that is closer to it, which results in the first-iteration clusters shown in Figure 1.



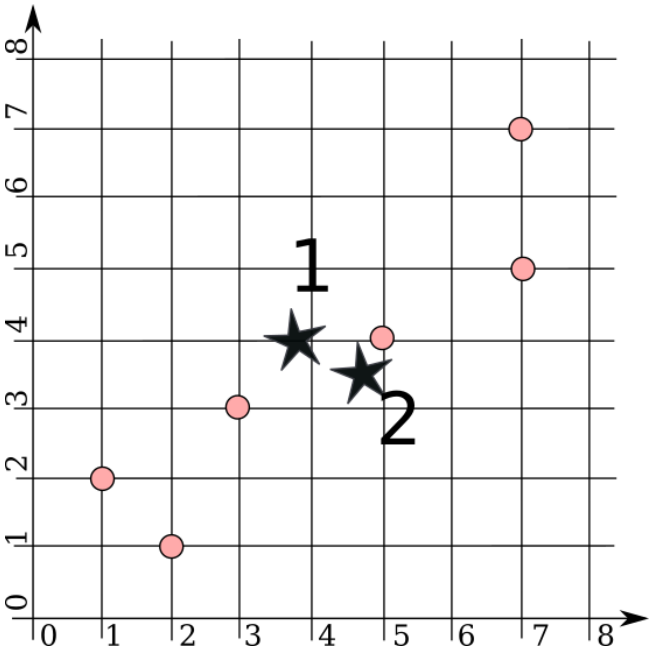
*****

Now we apply step 2, calculating the *centroids* of the current clusters.

Cluster 1 centroid: $x = \dfrac{1+3+7}{3} = 3.66, \quad y = \dfrac{2+3+7}{3} = 4$

Cluster 2 centroid: $x = \dfrac{2+5+7}{3} = 4.66, \quad y = \dfrac{1+4+5}{3} = 3.33$

We now use the calculated centroids as the cluster representative points. We remove the associations in the picture, as new ones may need to be made.



*****

What follows is the second application of step 1. For the points where the difference in distance is not immediately obvious, we calculate the distances.

For point $(2, 1)$ distance from the representative point of cluster 1 is:

$\sqrt{(3.66-2)^2 + (4-1)^2} = \sqrt{1.66^2 + 3^2} = 3.43$

The distance of this point from the representative point of cluster 2 is:

$\sqrt{(4.66-2)^2 + (3.33-1)^2} = \sqrt{2.66^2 + 2.33^2} = 3.54$

So we find **it should be associated with cluster 1**.

For point $(7, 7)$ distance from the representative point of cluster 1 is:

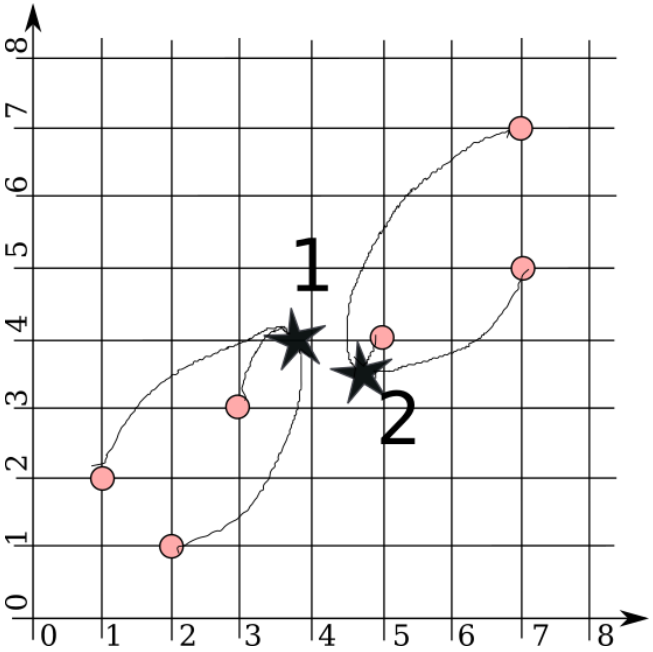
$\sqrt{(3.66-7)^2 + (4-7)^2} = 4.48$

The distance of this point from the representative point of cluster 2 is:

$\sqrt{(4.66-7)^2 + (3.33-7)^2} = 4.35$

So we find **it should be associated with cluster 2**.

For the other points the closer cluster-representative point is obvious from the picture.



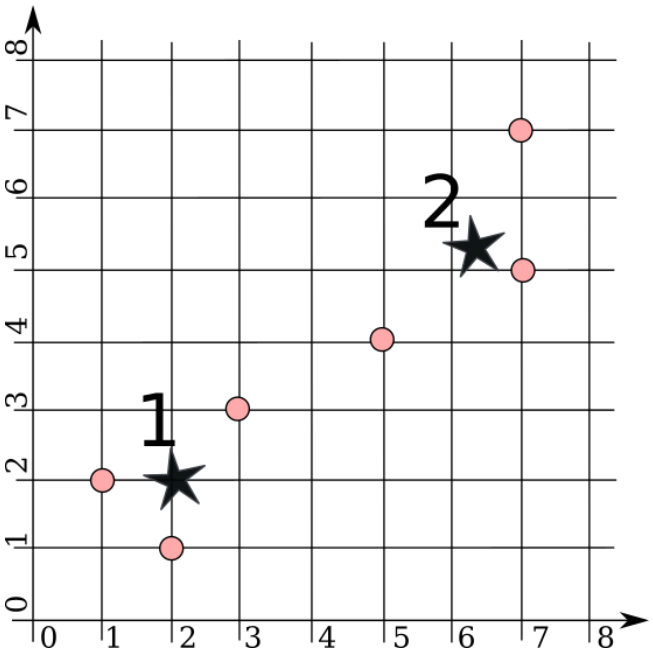
*****

We apply step 2 again, calculating the *centroids* of the current clusters.

Cluster 1 centroid: $x = \dfrac{1+2+3}{3} = 2, \quad y = \dfrac{1+2+3}{3} = 2$

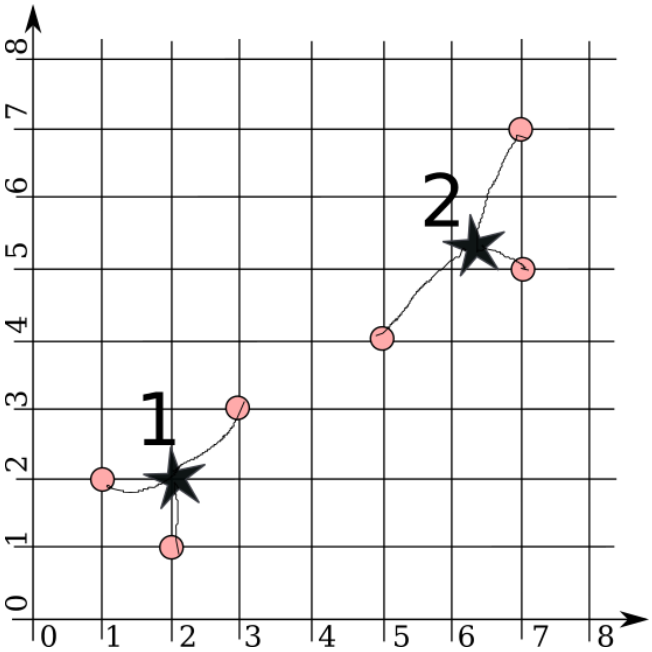
Cluster 2 centroid: $x = \dfrac{5+7+7}{3} = 6.33, \quad y = \dfrac{4+5+7}{3} = 5.33$

We use the calculated centroids as the new cluster-representative points.



*****

Step 1 is applied again. This time none of the points changes its association i.e. they all stay in the same clusters, which means that the process is finished and the clusters have been formed.



*****

ii Final cluster centroids and cluster membership by data instances

Cluster 1 centroid: (2, 2)

Cluster 2 centroid: (6.33, 5.33)

Points in cluster 1 are: $(2,1),\ (1,2)\ (3,3)$

Points in cluster 2 are: $(5,4),\ (7,5)\ (7,7)$

iii Applying the Elbow criterion for choosing the number of clusters

The within-cluster variance for the final clusters with $k = 2$ is:

$$WCV(2) = \frac{(1^2+0^2)+(0^2+1^2)+(1^2+1^2)+\left(\frac{4}{3}^2+\frac{4}{3}^2\right)+\left(\frac{2}{3}^2+\frac{1}{3}^2\right)+\left(\frac{2}{3}^2+\frac{5}{3}^2\right)}{6} = 1.89$$

Now we calculate the ratios of within-cluster variance for different values of k.

k	**WCV(k)**	**WCV(k)/WCV(k-1)**
1	9.36	NA
2	1.89	0.20
3	0.92	0.49
4	0.58	0.63

We are looking for the value of k for which $WCV(k)/WCV(k-1) < 0.5 < WCV(k+1)/WCV(k)$ i.e. the highest value of k for which the ratio of WCV to the WCV for one less cluster is below 0.5. By looking up the table we see that this is $k = 3$, with the ratio of 0.49, while the ratio for $k = 4$, being 0.63, is above 0.5.

Exercise 2 - k-Means Clustering ANSWER

Algorithm description:

The application of the k-means algorithm for clustering begins with the random choice of K data points to serve as initial cluster representative points. In the course of algorithm application these representative points move until they become centroids of stable (with respect to the algorithm) clusters that are the target and output of the algorithm. The choice of initial cluster representative points, however, has been completed in the question (C1 and C2 are given).

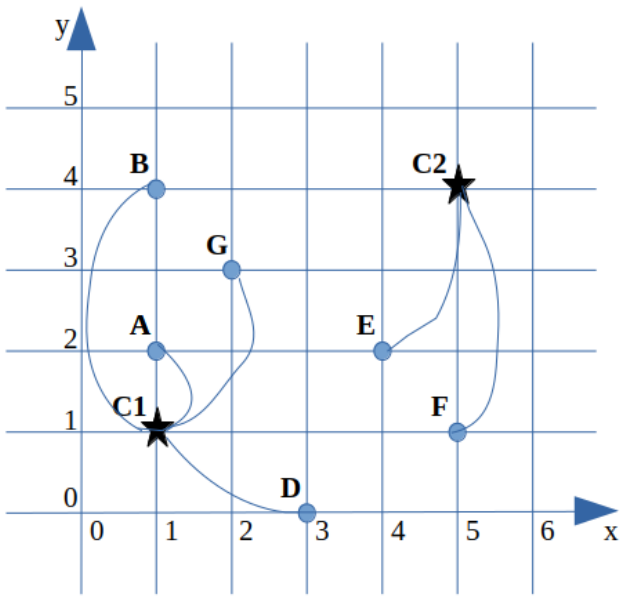
The algorithm application then consists of two steps that are performed in alteration, modifying the clusters until they stabilise. The clusters have stabilised once they haven't been modified by a pair of steps performed in sequence. The two steps are:

1. For each data point in the set, identify the cluster representative point closest in terms of Euclidean distance and attach the data point to the corresponding cluster.
2. Each cluster representative point is moved to assume the position of the centroid of its attached data points (at the arithmetic average for each of the dimensions, for all the attached data points).

Algorithm application:

Step 1: Closest cluster representative points for data points:

A: C1, B: C1, G: C1, D: C1, E: C2, F: C2

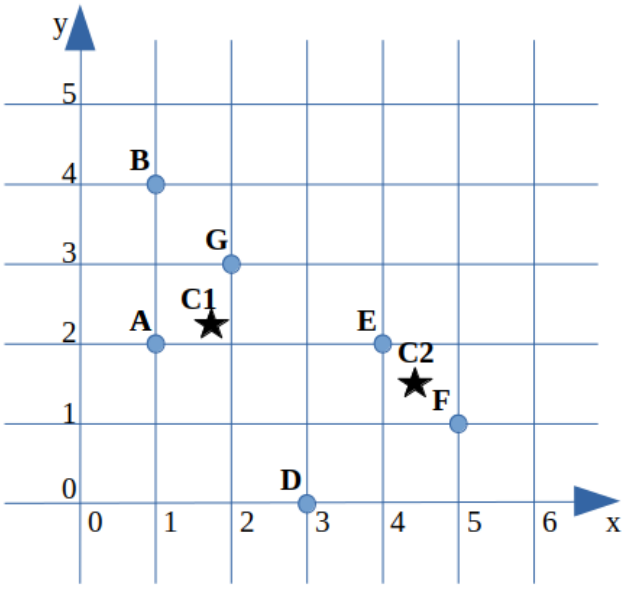


*****

Step 2: Move C1 and C2 so they assume centroid positions for the data points attached to them. We calculate and sketch the centroids:

C1: $x = \dfrac{1+1+2+3}{4} = 1.75,\ y = \dfrac{0+2+3+4}{4} = 2.25$

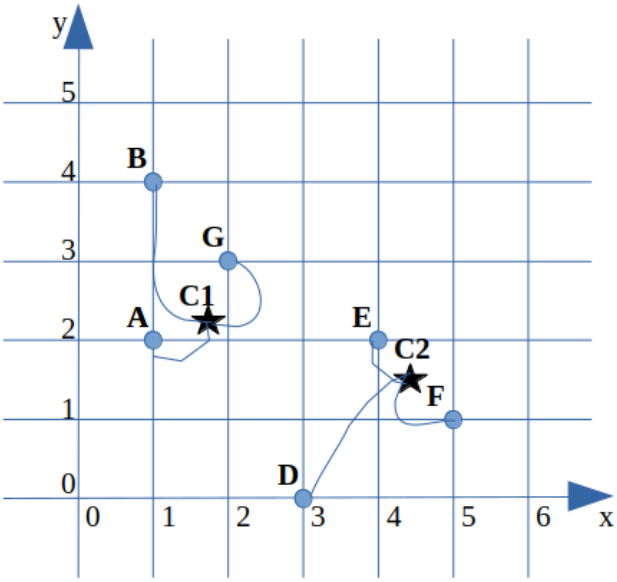
C2: $x = \dfrac{4+5}{2} = 4.5,\ y = \dfrac{1+2}{2} = 1.5$



*****

Step 1: Closest cluster representative points for data points:

A: C1, B: C1, G: C1, D: C2, E: C2, F: C2

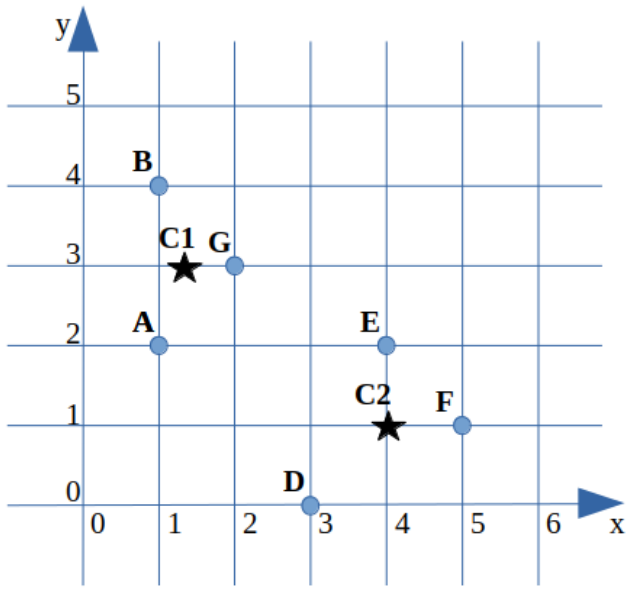


*****

Step 2: Move C1 and C2 so they assume centroid positions for the data points attached to them. We calculate and sketch the centroids:

C1: $x = \dfrac{1+1+2}{3} = 1.33,\ y = \dfrac{2+3+4}{3} = 3$

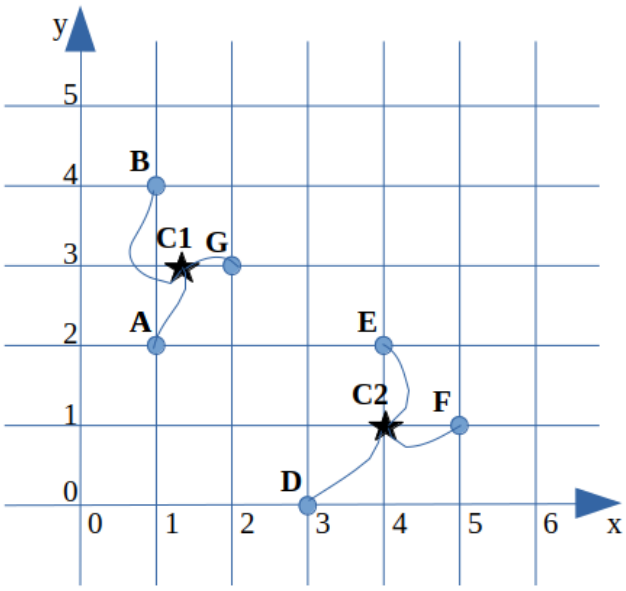
C2: $x = \dfrac{3+4+5}{3} = 4,\ y = \dfrac{0+1+2}{3} = 1$



*****

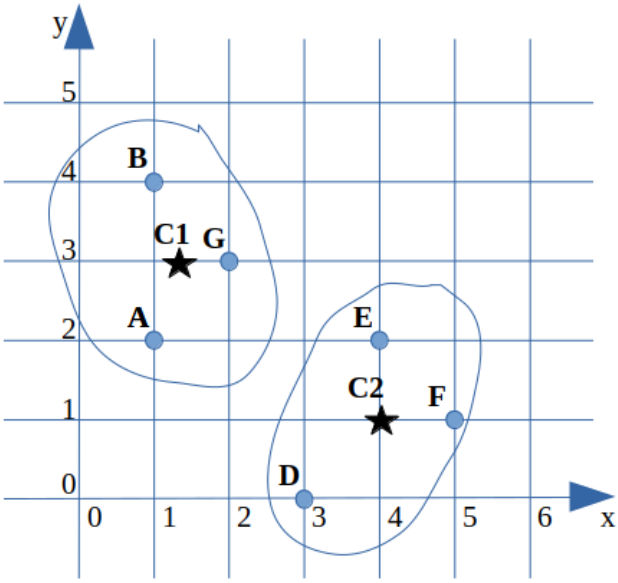
Step 1: Closest cluster representative points for data points:

A: C1, B: C1, G: C1, D: C2, E: C2, F: C2



*****

We note that in the last assignment of points to clusters no points have moved from one cluster to another i.e. the assignment was the same in the last two applications of step 1. This means that the clusters have become stable and can be pronounced defined.



Assignment of instances to clusters:

Cluster 1: A, B, G

Cluster 2: D, E, F

Cluster centroids:

C1: (1.33, 3)

C2: (4, 1)