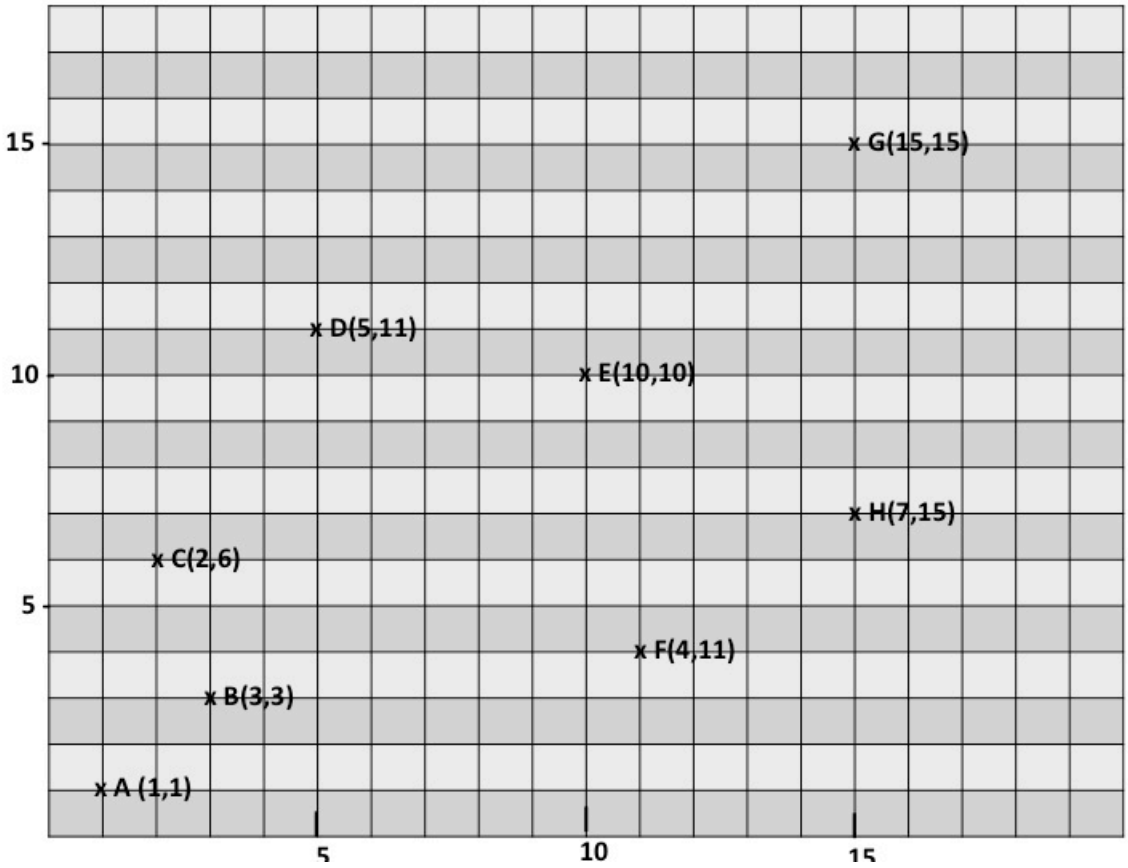
1. Consider the following 8 data objects, shown as 2-D points on a grid. Kmeans clustering is applied to find 3 clusters, using Euclidean distance. (Typos in the jpeg below: F should be (11,4) and H should be (15,7))



(a) (1.5 point) Suppose in the first iteration of Kmeans, the 3 points C, G, H are chosen as centroids to form the initial clusters. What are the clusters corresponding to C, G and H?

For C: {A, B, C, D}
For G: {G}
For H: {E, F, H}

(b) (1.5 points) For the initial clusters in (a), what are the new centroids and new clusters at the end of the second iteration of Kmeans?

For C: (2.75,5.25)
For G: (15,15)
For H: (7,12)
No Change as in (a).

(c) (2 points) Consider the silhouette coefficient from the first assignment:

$$s(x_i) = [b(x_i) - a(x_i)] / \max\{a(x_i), b(x_i)\}$$

where

- $a(x_i)$ is the average dissimilarity between x_i and all other objects in the cluster to which x_i belongs; and
- $b(x_i)$ is the average dissimilarity between x_i and all objects in the nearest cluster to which x_i does not belong.

Using the clusters in part (a), identify the data object with the worst silhouette coefficient and the one with the best silhouette coefficient. You do not need to show your calculations.

Worst: D
Best: G