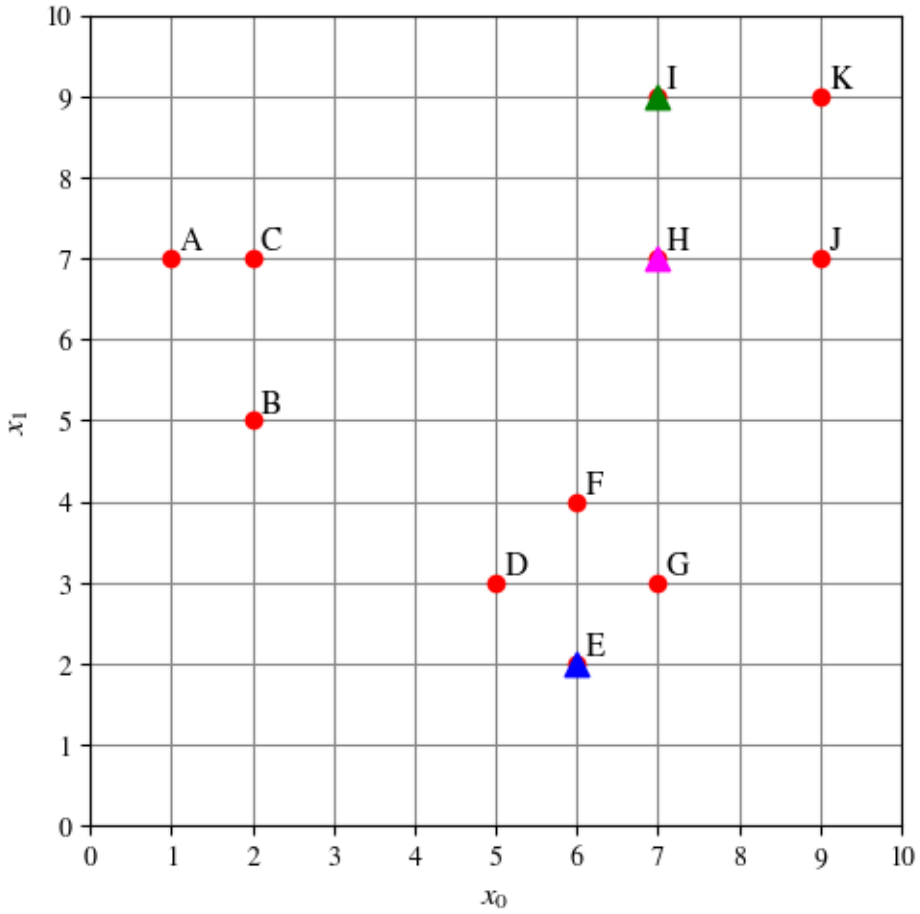
7 Unsupervised Learning

7.1 K-Means Clustering



i. In the figure above, the scatterplot of the dataset with the 3 initial cluster centers is shown. The clusters will be referred as cluster-1 (green centroid), cluster-2 (pink centroid) and cluster-3 (blue centroid).

   By visual inspection we can assign clusters to each data sample as follows:

   - cluster-1: Points K and I.
   - cluster-2: Points A, C, H and J.
   - cluster-3: Points B, D, E, F and G.

ii. The new cluster centers can be obtained by calculating the centroid of the data samples assigned to each cluster in the previous subtask.

- cluster-1:

$$K = \begin{bmatrix} 9 \\ 9 \end{bmatrix}, \quad I = \begin{bmatrix} 7 \\ 9 \end{bmatrix}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} \frac{9+7}{2} \\ \frac{9+9}{2} \end{bmatrix} = \begin{bmatrix} 8 \\ 9 \end{bmatrix}$$

- cluster-2:

$$A = \begin{bmatrix} 1 \\ 7 \end{bmatrix}, \quad C = \begin{bmatrix} 2 \\ 7 \end{bmatrix}, \quad H = \begin{bmatrix} 7 \\ 7 \end{bmatrix}, \quad J = \begin{bmatrix} 9 \\ 7 \end{bmatrix}$$

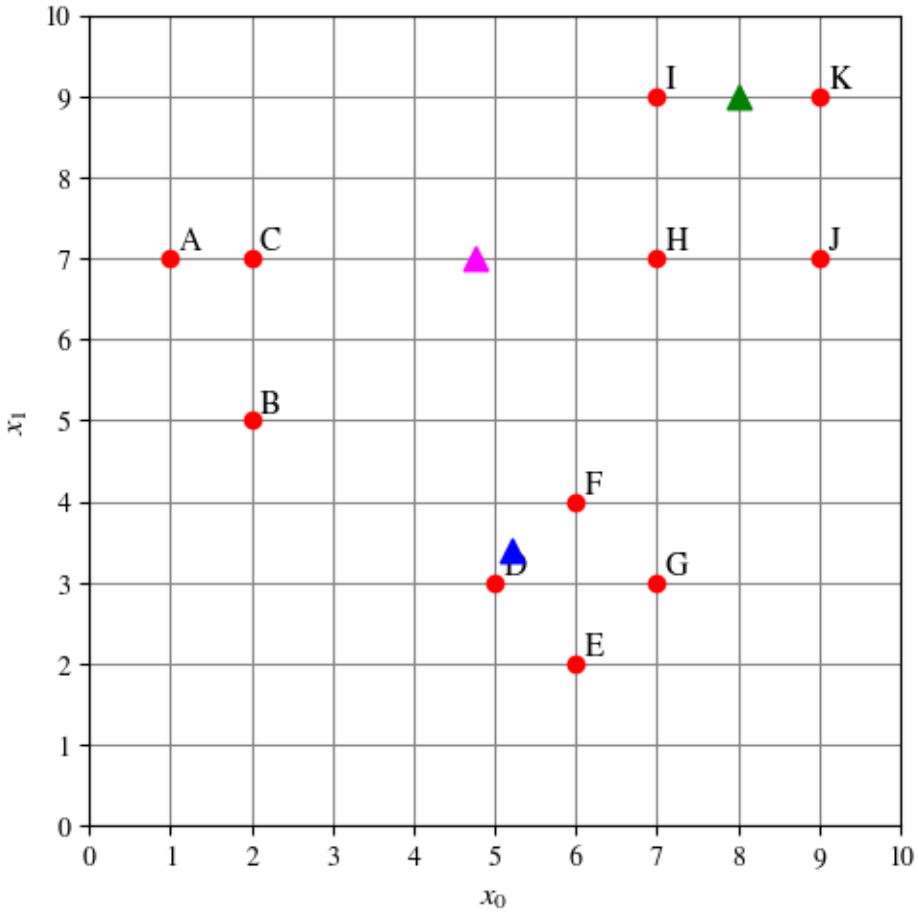
$$\boldsymbol{\mu}_2 = \begin{bmatrix} \frac{1+2+7+9}{4} \\ \frac{7+7+7+7}{4} \end{bmatrix} = \begin{bmatrix} 4.75 \\ 7 \end{bmatrix}$$

- cluster-3:

$$B = \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \quad D = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \quad E = \begin{bmatrix} 6 \\ 2 \end{bmatrix}, \quad F = \begin{bmatrix} 7 \\ 3 \end{bmatrix}, \quad G = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} \frac{2+5+6+7+6}{5} \\ \frac{5+3+2+3+4}{5} \end{bmatrix} = \begin{bmatrix} 5.2 \\ 3.4 \end{bmatrix}$$

NOTE: In order to calculate the objective function after the first iteration [subtask (iii)], we need the cluster assignments for the new cluster centers. Hence we solve subtask (iv) before subtask (iii).



iv. In the figure above, the scatterplot of the dataset with the new cluster centers after the first iteration is shown.

By visual inspection we can assign clusters to each data sample as follows:

- cluster-1: Points H, I, J and K.
- cluster-2: Points A, B and C.
- cluster-3: Points D, E, F and G.

iii. The objective function of the K-Means algorithm for K clusters where N_k is number of samples in a cluster k, is given by:

$$\begin{aligned}\mathcal{L} &= \sum_{k=1}^{K}\sum_{i=1}^{N_k}\|\boldsymbol{\mu}_k - \mathbf{x}_i\|^2 \\ &= \quad \|\boldsymbol{\mu}_1 - H\|^2 + \|\boldsymbol{\mu}_1 - I\|^2 + \|\boldsymbol{\mu}_1 - J\|^2 + \|\boldsymbol{\mu}_1 - K\|^2 \\ &+ \quad \|\boldsymbol{\mu}_2 - A\|^2 + \|\boldsymbol{\mu}_2 - B\|^2 + \|\boldsymbol{\mu}_2 - C\|^2 \\ &+ \quad \|\boldsymbol{\mu}_3 - D\|^2 + \|\boldsymbol{\mu}_3 - E\|^2 + \|\boldsymbol{\mu}_3 - F\|^2 + \|\boldsymbol{\mu}_3 - G\|^2 \\ &= 52.39\end{aligned}$$