COEN 240 Machine Learning

Assignment # 2

This assignment is about allocating the data points to three clusters.

Input

The data-file is:

$$Data\ Points.xlxs \qquad (EXCEL\ file)$$

Note that in the Excel Work Sheet *Data Points*: Column A has the data label, and columns B and C have the x and y coordinates. There are a total of 32 data points. In the file

$$Assignment\ \#2\ Plot\ of\ Data\text{-}Points.pdf$$

observe three distinct clusters in the plot of the data points in the x-y plane.

Algorithm

Using the K-Means clustering algorithm, allocate the data points to the three clusters.

1. Use a suitable stopping criteria for stopping the algorithm.
2. The proximity measure for the data points is the Euclidean distance. Note that the Euclidean distance between the points

   $$z_1 = (x_1, y_1) \quad \text{and} \quad z_2 = (x_2, y_2)$$

   is

   $$d(z_1, z_2) = \left\{ (x_1 - x_2)^2 + (y_1 - y_2)^2 \right\}^{1/2}$$

3. The centroid of points in the set $\{(x_i, y_i) \mid 1 \leq i \leq n\}$ is $(\overline{x}, \overline{y})$, where

   $$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Expected Output

Denote the data label set by L. It is

$$L = \{1, 2, 3, \ldots, 32\}$$

1. List of the initial three centroids (coordinates). These centroids are not necessarily the coordinates of any of the data points.
2. Let the three clusters be C_1, C_2, and C_3. Output the data clusters as:

(a) $C_1 = \{a, b, \ldots, u\}$, $|C_1|$, and centroid (X_1, Y_1) at the end of the algorithm.

(b) $C_2 = \{c, d, \ldots, v\}$, $|C_2|$, and centroid (X_2, Y_2) at the end of the algorithm.

(c) $C_3 = \{e, f, \ldots, w\}$, $|C_3|$, and centroid (X_3, Y_3) at the end of the algorithm.

where $a, b, c, d, e, f, u, v, w \in L$.

3. Check that
$$|C_1| + |C_2| + |C_3| = 32$$

4. Number of iterations that were required for convergence.

You should submit the high-level code and the output in a pdf file. Please label the file with your name on it.