CSE708: Application of Data analytics and Engineering

Fall 2023

Assignment2: K-means Clustering

Due on Friday September 29, 2023, at 11:59PM

This assignment must be completed by hand calculations. Please don't use any programming language.

I would like to put eight data points into 3 different groups with groups: A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9).
Find the 3 groups using k-means algorithm after 2 iterations.
(**Instructions**: Consider C1=A, C2=D and C3=G as initial centers. And you must use Manhattan distance.)

Hint 1. Fill out this table and decide based on your results,
Hint 2. Recompute the new centers,
Hint 3. Repeat hint 1 basing on new centers.

Data Points	Distance from C1	Distance from C2	Distance from C3	It belongs to the cluster
A(2,10)				
B(2,5)				
C(8,4)				
D(5,8)				
E(7,5)				
F(6,4)				
G(1,2)				
H(4,9)				

Good Luck!

Name: Robert Akinie

Assignment 2

CSE 708

Given the following data points above, and initial centroids, the distance between the individual points and the centroids are computed using the Manhattan distance, given below as:

$$d(x, y) = | x_1 - x_2 | + | y_1 - y_2 |$$

The first iteration has results below

Data Points	Distance from C1	Distance from C2	Distance from C3	It belongs to the cluster
A(2, 10)	|2-2|+|10-10| = 0	|5-2|+|8-10| = 5	|1-2|+|2-10| = 9	C1
B(2, 5)	|2-2|+|10-5| = 5	|5-2|+|8-5| = 6	|1-2|+|2-5| = 4	C3
C(8, 4)	|2-8|+|10-4| = 12	|5-8|+|8-4| = 7	|1-8|+|2-4| = 9	C2
D(5, 8)	|2-5|+|10-8| = 5	|5-5|+|8-8| = 0	|1-5|+|2-8| = 7	C2
E(7, 5)	|2-7|+|10-5| = 10	|5-7|+|8-5| = 5	|1-7|+|2-5| = 9	C2
F(6, 4)	|2-6|+|10-4| = 10	|5-6|+|8-4| = 5	|1-6|+|2-4| = 7	C2
G(1, 2)	|2-1|+|10-2| = 9	|5-1|+|8-2| = 10	|1-1|+|2-2| = 0	C3
H(4, 9)	|2-4|+|10-9| = 3	|5-4|+|8-9| = 2	|1-4|+|2-9| = 10	C2

Following this results, the new centroids are computed, based on determining the average between the points in the clusters found.
The average is given as:

$$\text{avg}(x, y) = \left(\frac{x_1 + x_2 + \ldots + x_n}{n} , \frac{y_1 + y_2 + \ldots + y_n}{n} \right)$$

Using the average, the new centroids are as follows:
C1(x, y) = (2, 10)

$$\text{C2}(x, y) = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6)$$

$$\text{C3}(x, y) = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

From these new computed centroids, the second iteration is taken to compute the new clusters as follows:

Data Points	Distance from C1 (2, 10)	Distance from C2 (6, 6)	Distance from C3 (1.5, 3.5)	It belongs to the cluster
A(2, 10)	|2-2|+|10-10| = 0	|6-2|+|6-10| = 8	|1.5-2|+|3.5-10| = 7	C1
B(2, 5)	|2-2|+|10-5| = 5	|6-2|+|6-5| = 3	|1.5-2|+|3.5-5| = 2	C3
C(8, 4)	|2-8|+|10-4| = 12	|6-8|+|6-4| = 4	|1.5-8|+|3.5-4| = 7	C2
D(5, 8)	|2-5|+|10-8| = 5	|6-5|+|6-8| = 3	|1.5-5|+|3.5-8| = 8	C2
E(7, 5)	|2-7|+|10-5| = 10	|6-7|+|6-5| = 2	|1.5-7|+|3.5-5| = 7	C2
F(6, 4)	|2-6|+|10-4| = 10	|6-6|+|6-4| = 2	|1.5-6|+|3.5-4| = 5	C2
G(1, 2)	|2-1|+|10-2| = 9	|6-1|+|6-2| = 9	|1.5-1|+|3.5-2| = 2	C3
H(4, 9)	|2-4|+|10-9| = 3	|6-4|+|6-9| = 5	|1.5-4|+|3.5-9| = 8	C1

Following this results, the new centroids are computed, based on determining the average between the points in the clusters found.

Using the average, the new centroids are as follows:

C1(x, y) = $\left(\frac{2+4}{2}, \frac{10+9}{2}\right)$ = (3, 9.5)

C2(x, y) = $\left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right)$ = (6.5, 5.25)

C3(x, y) = $\left(\frac{2+1}{4}, \frac{5+2}{4}\right)$ = (1.5, 3.5)

The new centers from the second iteration are (3, 9.5), (6.5, 5.25) and (1.5, 3.5)