[Subratim Nag - CSE-AIML/22157 - ML Assignment - CA3]

Q1 Lower dimensional data can be represented in higher dimensions using various techniques, often involving transformations or mapping that preserve certain properties of the original data.

example:-

Consider a 2-D datasets consisting of points in a plane: $D_2 = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

each point represent 2-D co-ordinate,

in 3D:-

now to represent in 3D, have to map each 2D point (x_i, y_i) to a 3D point $(x_i, y_i, f(x_i, y_i))$, where $f(x_i, y_i)$ is a function that generates the third co-ordinate based on some criteria,

let use, $f(x_i, y_i) = x_i^2 + y_i^2$

the function $f(x_i, y_i)$ maps each 2D points to a 3D point where the third co-ordinate is based on squared Euclidean distance,

thus, our dataset D becomes (3D):

$$D_3 = \{(x_1, y_1, x_1^2+y_1^2), (x_2, y_2, x_2^2+y_2^2), \ldots (x_n, y_n, x_n^2+y_n^2)\}$$

$\Rightarrow$ In 2D, points are spread out in a plane,

In 3D, same points are distributed on a curve surface, reflecting their original 2D positions.

2) Given Dataset, $D = \{2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30\}$

Cluster centers, $C_1(2)$, $C_2(16)$, $C_3(38)$.

Steps:-

i) Start with initial cluster center

ii) Assign each data point to the nearest cluster center

iii) Calculate new cluster centers as the mean of all points

iv) repeat step ii & iii, until cluster assignments do not change

Iteration 1:-

Data Point	Distance to $C_1(2)$	Distance to $C_2(16)$	Distance to $C_3(38)$	Nearest Cluster
2	0	14	36	C_1
4	2	12	34	C_1
6	4	10	32	C_1
3	1	13	35	C_1
31	29	15	7	C_3
12	10	4	26	C_2
15	13	1	23	C_2
16	14	0	22	C_2
38	36	22	0	C_3
35	33	19	3	C_3
14	12	2	24	C_2
21	19	5	17	C_2
23	21	7	15	C_2
25	23	9	13	C_2
30	28	14	8	C_3

Data Point	Distance to $C_1(2)$	Distance to $C_2(16)$	Distance to $C_3(38)$	Nearest Cluster
25	23	9	13	C_2
30	28	14	8	C_3

Updating the clusters,

$$C_1 = \frac{2+4+6+3}{4} = 3.75$$

$$C_2 = \frac{12+15+16+14+21+23+25}{7} = \frac{126}{7} = 18$$

$$C_3 = \frac{31+38+35+30}{4} = \frac{134}{4} = 33.5$$

Iteration 2:-

Data Point	Distance to $C_1(3.75)$	Distance to $C_2(18)$	Distance to $C_3(33.5)$	Nearest Cluster
2	1.75	16	31.5	C_1
4	0.25	14	29.5	C_1
6	2.25	12	27.5	C_1
3	0.75	15	30.5	C_1
31	27.25	13	2.5	C_3
12	8.25	6	21.5	C_2
15	11.25	3	18.5	C_2
16	12.25	2	17.5	C_2
38	34.25	20	4.5	C_3
35	31.25	17	1.5	C_3
14	10.25	4	19.5	C_2
21	17.25	3	12.5	C_2
23	19.25	5	10.5	C_2
25	21.25	7	8.5	C_2
30	26.25	12	3.5	C_3

update clusters:-

$$c_1 = \frac{2+4+6+3}{4} = 3.75$$

$$c_2 = \frac{12+15+16+14+21+23+25}{7} = 18$$

$$c_3 = \frac{31+38+35+30}{4} = 33.5$$

Convergence:-

Since, cluster centers have not changed in this iteration, k-means algo has converged.

$c_1 = \{2, 4, 6, 3\}$, $c_2 = \{12, 15, 16, 14, 21, 23, 25\}$

$c_3 = \{31, 38, 35, 30\}$

So, best cluster would be,

$c_1 = \{2, 4, 6, 3\}$ with center 3.75

$c_2 = \{12, 15, 16, 14, 21, 23, 25\}$ with center 18

$c_3 = \{31, 38, 35, 30\}$ with center 33.5

Ans