Name- Mudit Sanad

Roll No - 20310068

EE 769 - Tutorial 6

Ans + ① Loss function for K-means Clustering:-

$$J = \sum_{n=1}^{N} \sum_{n=1}^{K} \gamma_{nk} \|x_n - \mu_k\|^2$$

where $\gamma_{nk} = \begin{cases} 1 & \text{if } k = \arg\min \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$

From the Hint 1) given in the Question

1) Centroid Location minimizes Sum of Squaredistances,

Consider $\vec{x}_1 - \vec{x}_n$

Minimize distance to some point $\vec{t}$

$\min \sum_i |\vec{x}_i - \vec{t}|^2$ If c is the Centroid.

$= \min \sum_i |\vec{x}_i - \vec{c} + \vec{c} - \vec{t}|^2$

$= \min \sum_i |\vec{x}_i - \vec{c}|^2 + 2(c-t)\sum|x_i - c| + n|\vec{c} - \vec{t}|^2$

If c is centroid then $\sum(\vec{x}_i - \vec{c}_i) = 0$

So $\Rightarrow \min \left[\sum_i |\vec{x}_i - \vec{c}|^2 + n|\vec{c} - \vec{t}|^2\right]$

To minimize the above expression we can

Choose our Centroid a t itself.

That means $|\vec{c} - \vec{t}| = 0$

Ans

$$\Rightarrow \min \sum_{i} |\vec{x}_i - \vec{c}_i|^2$$

from hint 2)

1) Cluster assignment to nearest center will minimize sum of square distances

and Above two steps ensure that between two iteration cost cannot increase..

Answer → ② So the given set of points mentioned below

$A = (0,0)^T$ $B = (0,1)^T$ $C = (1,1)^T$, $D = (3,3)^T$ $E = (3,4)^T$
$F = (2.5, 3.5)^T$

given cluster centers are $P = (-1,5)^T$, $Q = (5,0)^T$

Ist Iteration:

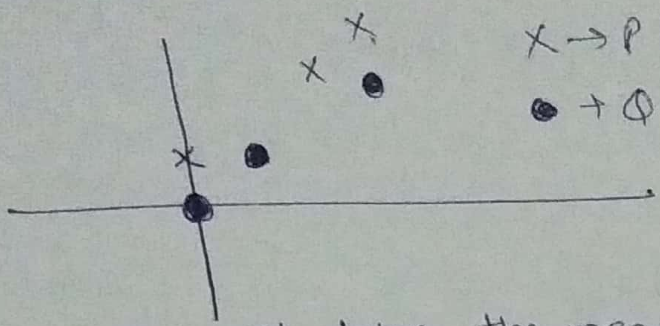
Distance b/w various points with two cluster centers are

	A	B	C	D	E	F
P	5.09	4.12	4.47	4.47	4.12	8.8
Q	5	5.09	4.12	3.6	4.24	4.30

Point	A	B	C	D	G	F
Cluster	Q	P	Q	Q	P	P

{from Minm Distance by above given table}

If we visualize this ↓



Now we will take the mean and update the Center value

II nd Iteration:

$$p' = \left(\frac{0+3+2.5}{3}, \frac{1+4+3.5}{3} \right)$$

$$p' = (1.833, 2.833) \qquad q' = \left(\frac{0+1+3}{3}, \frac{0+1+3}{3} \right)$$

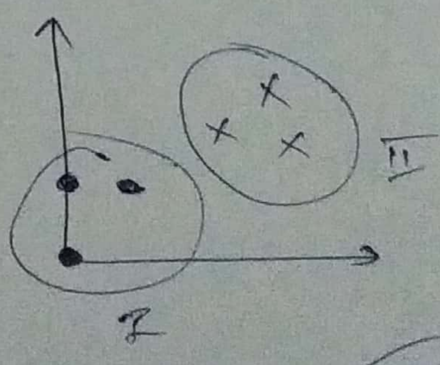
$$q' = (1.33, 1.33)$$

Now if we Calculate the distance from New Cluster Center than

	A	B	C	D	E	F
P	3.37	2.59	2.01	1.178	1.65	.943
Q	1.88	1.37	.466	2.361	3.145	2.46

Point	A	B	C	D	E	F
Cluster	Q	Q	Q	P	P	P

So it is Clearly visible that in two iteration two Clusters are dividing these Points.



Ans → 3 Part a) Single Linkage (min. distance)

Points → A (0,0), B (2,0), C (5,0), D (2,3), E (0,6), F (0,8)
G (0,11)

Matrix → (cell Represent Distance)

	A	B	C	D	E	F	G
A	0	2	5	$\sqrt{13}$	6	8	11
B	2	0	3	3	$\sqrt{40}$	$\sqrt{68}$	$\sqrt{125}$
C	5	3	0	$\sqrt{18}$	$\sqrt{61}$	$\sqrt{89}$	$\sqrt{146}$
D	$\sqrt{13}$	3	$\sqrt{18}$	0	$\sqrt{13}$	$\sqrt{29}$	$\sqrt{68}$
E	6	$\sqrt{40}$	$\sqrt{61}$	$\sqrt{13}$	0	2	5
F	8	$\sqrt{68}$	$\sqrt{89}$	$\sqrt{29}$	2	0	3
G	11	$\sqrt{125}$	$\sqrt{146}$	$\sqrt{68}$	5	3	0

So Minimum distance is for E & F ⇒ 2
Now New Matrix. for Single Linkage (min. distance)
We will Carry $d = \min \begin{Bmatrix} (\text{Point}) E \\ (\text{Point}) F \end{Bmatrix}$ in matrix

	A	B	C	D	$\boxed{EF}$	G
A	0	2	5	$\sqrt{13}$	6	11
B		0	3	3	$\sqrt{40}$	$\sqrt{125}$
C			0	$\sqrt{18}$	$\sqrt{61}$	$\sqrt{146}$
D				0	$\sqrt{13}$	$\sqrt{68}$
$\boxed{EF}$					0	3
G						0

Scanned with CamScanner

Now Min^m for A & B

	AB	C	D	EF	G
AB	0	3	3	6	11
C		0	$\sqrt{18}$	$\sqrt{61}$	$\sqrt{146}$
D			0	$\sqrt{13}$	$\sqrt{68}$
EF				0	3
G					0

Now min is 3

we will take AB C

	ABC	D	EF	G
ABC	0	3	6	11
D		0	$\sqrt{13}$	$\sqrt{68}$
EF			0	3
G				0

Now min is 3

take ABCD

	ABCD	EF	G
ABCD	0	$\sqrt{13}$	$\sqrt{68}$
EF		0	3
G			0

Min is 3

So EF G

final Clusters for Single Linkage (Min. Distance)

{A, B, C, D} and {E, F, G}

I II

Part (b) Complete Linkage (Max distance)

From the first matrix we can see the max distance is for GC is $\sqrt{146}$

Constructing New matrix

	A	B	D	E	F	GC
A	0	2	$\sqrt{13}$	6	8	11
B		0	3	$\sqrt{40}$	$\sqrt{68}$	$\sqrt{125}$
D			0	$\sqrt{13}$	$\sqrt{29}$	$\sqrt{68}$
E				0	2	$\sqrt{61}$
F					0	$\sqrt{89}$
GC						0

→ Now max is for BGC

	A	D	E	F	BGC
A	0	$\sqrt{13}$	6	8	11
D		0	$\sqrt{13}$	$\sqrt{29}$	$\sqrt{68}$
E			0	2	$\sqrt{61}$
F				0	$\sqrt{89}$
BGC					0

Now Max is ABGC

	D	E	F	ABGC
D	0	$\sqrt{13}$	$\sqrt{29}$	$\sqrt{68}$
E		0	2	$\sqrt{61}$
F			0	$\sqrt{89}$
ABGC				0

Max is for FABGC

	D	E	FABGC
D	0	$\sqrt{13}$	$\sqrt{68}$
E		0	$\sqrt{61}$
FABGC			0

So looking at this for Complete linkage only I Cluster will be there (2) $\{A, B, C, D, E, F, G\}$

Part c we need to find Avg. distance b/w all pairs of points in two Cluster

$I \rightarrow A, B, C, D$ $\Rightarrow d_{Avg}[(AB), (AC), (AD)]$

$II \rightarrow E, F, G$ $\Rightarrow d_{avg}[2, 5, \sqrt{13}]$

$\Rightarrow d_{avg_I}(A) = 3.535$

$d_{avg}[(EF), (EG)]$

$\Rightarrow d_{avg}[2, 5] = 3.5$

So one can see avg. distance is almost same.

Ans - 4 PCA on $\{[0,0]^T, [6,6]^T, [0,1]^T, [6,7]^T\}$ (8)

(a) Centroid of the data :-

$$x = \begin{bmatrix} 0 & 0 \\ 6 & 6 \\ 0 & 1 \\ 6 & 7 \end{bmatrix} \qquad \text{Centroid} = \begin{bmatrix} 3 & 3.5 \end{bmatrix}$$

Subtracting the data from Centroid

$$x' = \begin{bmatrix} -3 & -3.5 \\ 3 & 2.5 \\ -3 & -2.5 \\ 3 & 3.5 \end{bmatrix}$$

(b) Computing the Covariance matrix C

$$C = [x']^T x'$$

$$= \begin{bmatrix} -3 & 3 & -3 & 3 \\ -3.5 & 2.5 & -2.5 & 3.5 \end{bmatrix} \begin{bmatrix} -3 & -3.5 \\ 3 & 2.5 \\ -3 & -2.5 \\ 3 & 3.5 \end{bmatrix}$$

$$C = \begin{bmatrix} 36 & 36 \\ 36 & 37 \end{bmatrix}$$

Now to Calculate the Eigen Values and Vectors we need to solve $|C - \lambda I| = 0$

$$\begin{vmatrix} 36-\lambda & 36 \\ 36 & 37-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (36-\lambda)(37-\lambda) - 36^2 = 0$$

$\lambda = .4965 \,,\, 72.5035$
$\quad\ \downarrow \qquad\quad \downarrow$
$\quad\ \lambda_1 \qquad\quad \lambda_2$

Eigen vector corresponding to λ_1:

$$(C - \lambda_1 I)\, v_1 = 0$$

$$\begin{bmatrix} 36 - .4965 & 36 \\ 36 & 37 - .4965 \end{bmatrix} (v_1) = 0$$

$$v_1 = \begin{bmatrix} -.7120 \\ .7022 \end{bmatrix}$$

Eigen vector corresponding to λ_2

$$[C - \lambda_2 I]\, v_2 = 0$$

$$\begin{bmatrix} 36 - 72.5035 & 36 \\ 36 & 37 - 72.5035 \end{bmatrix} (v_2) = 0$$

$$\Rightarrow v_2 = \begin{bmatrix} .7022 \\ .7120 \end{bmatrix}$$

(c) Confirming orthogonality

$\Rightarrow v_1^T v_2$

$\Rightarrow \begin{bmatrix} -.7120 & .7022 \end{bmatrix} \begin{bmatrix} .7022 \\ .7120 \end{bmatrix}$

$\Rightarrow 0$ {Hence proved this is orthogonal}

Part (d) Finding the transformed Coordinate L

$T = x' \times [v_1, v_2]$

$$T = \begin{bmatrix} -.3216 & -4.5985 \\ -.3806 & 3.8865 \\ .3806 & -3.8865 \\ .3216 & 4.5985 \end{bmatrix}$$

Part (e) when we Calculate the Variance of first Column is transformed Coordinate MatrixL i.e. ⇒ 0.1655

Variance of Second Column ⇒ 24.168

$Var_2 >>> Var_1$

So we are taking the information with minimal loss in variance by Choosing Column :- [-4.5985 3.8865 -3.8865 4.5985]