Assignment 4(CLL788)

The code for this assignment is uploaded as python notebook file CLL788_Assignment4.ipynb.

2 (e). On comparing the kind of results which we get, we see that both K-Means and GMM give us quite accurate results. However, GMM is a better method as it can capture variance better. GMM can make elliptical decision boundaries whereas K-Means cannot do that. But in terms of computational speed, K-means is faster method because it requires less calculations as compared to GMM.

Name - Ambrish Kashyap
Entry - 2019CH70157

papergrid
Date: / /

CLL 788 Assignment - 4

(1)

S.No.	Var 1	Var 2
1	-1.54	2.29
2	-0.44	2.34
3	0.03	0.41
4	1.2	1.87
5	0.65	2.39
6	-4.67	-4.8
7	-3.37	-5.41
8	-3.93	-4.64
9	-4.78	-4.96
10	-4.12	-5.36

Table 1

We assume var 1 to be x & var 2 to be y
∴ Each point can be written as (x_i, y_i) $[1 \le i \le 10]$

∴ We have to divide these points into 2 clusters
∴ We assume 2 means : $\mu_1 (x_3, y_3)$ & $\mu_2 (x_6, y_6)$

S. No.	Distance from μ_2	Distance from μ_1	Cluster
1	7.75	2.44	1
2	8.29	1.98	1
3	7.01	0	1
4	8.88	1.87	1
5	8.94	2.07	1
6	0	7.01	0
7	1.43	6.74	0
8	0.75	6.41	0
9	0.19	7.20	0
10	0.78	7.1	0

Table 2

We calculate the distance of each point from both the assumed means μ_0 & μ_1 using the formula

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{(\mu x - x_i)^2 + (\mu y - y_i)^2}$$

We assign the cluster to each point on basis of its ~~Now, we assume~~ distance from the assumed means. ~~The mean of~~ Points were assigned to the cluster which had less distance from mean.

Now, we assume the cluster assignment as constant and would shift the mean.

∴ For cluster 0:

x-coordinate of new mean =

$$= \frac{-4.67 - 3.37 - 3.93 - 4.78 - 4.12}{5}$$

$$= -4.174$$

y-coordinate of new mean

$$= \frac{-4.8 - 5.41 - 4.64 - 4.96 - 5.36}{5}$$

$$= -5.034$$

Similarly, for cluster 1:

x-coordinate = $\frac{-1.54 - 0.44 + 0.03 + 1.2 + 0.65}{5} = -0.02$

y-coordinate = $\frac{2.29 + 2.34 + 0.41 + 1.87 + 2.39}{5} = 1.86$

⇒ $\mu_0 = (-4.174, -5.034)$ & $\mu_1 = (-0.02, 1.86)$

S.No.	Distance from μ_0	Distance from μ_1	Cluster
1	7.78	1.57	0 1
2	8.26	0.63	0 1
3	6.87	1.45	0 1
4	8.74	1.22	0 1
5	8.85	0.85	0 1
6	0.54	8.12	1 0
7	0.88	8.0	1 0
8	0.46	7.58	1 0
9	0.61	8.31	1 0
10	0.33	8.30	1 0

Table 3

Now, we shift the mean

For cluster 0: $X = \frac{-4.67 - 3.37 - 3.93 - 4.78 - 4.12}{5}$

$= -4.174$

$Y = \frac{-4.8 - 5.41 - 4.64 - 4.96 - 5.36}{5}$

$= -5.034$

For cluster 1: $X = \frac{-1.54 - 0.44 + 0.03 + 1.2 + 0.65}{5}$

$= -0.02$

$Y = \frac{2.29 + 2.34 + 0.41 + 1.87 + 2.39}{5}$

$= 1.86$

⇒ [$\mu_0 : (-4.174, -5.034)$ & $\mu_1 : (-0.02, 1.86)$]

∵ The means are remaining same

⇒ They have converged ⇒ Table 3 shows correct classification

(3)

Y1	Y2
2	1
3	4
5	0
7	6
9	2

We first make the data centred around 0 by subtracting the mean from both columns.

Mean for $Y1 = \frac{2+3+5+7+9}{5} = \frac{26}{5} = 5.2$

Mean for $Y2 = \frac{1+4+0+6+2}{5} = \frac{13}{5} = 2.6$

New Table :

Y1	-3.2	-2.2	-0.2	1.8	3.8
Y2	-1.6	1.4	-2.6	3.4	-0.6

Now, we calculate the covariance matrix S.

∴ We have 20 data

∴ S has dimensions 2×2

$$S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

~~$S_{11} = \frac{1}{5}$~~ We have already made x values to be $x - \bar{x}$

$$\Rightarrow S_{11} = \frac{1}{5}\left[(-3.2)^2 + (-2.2)^2 + (-0.2)^2 + (1.8)^2 + (3.8)^2\right]$$

$$= \frac{1}{5}\left[10.24 + 4.84 + 0.04 + 3.24 + 14.44\right]$$

$$= 6.56$$

$$S_{22} = \frac{1}{5}\left[(-1.6)^2 + (1.4)^2 + (-2.6)^2 + (3.4)^2 + (-0.6)^2\right]$$

$$= \frac{1}{5}[2.56 + 1.96 + 6.876 + 11.56 + 0.36]$$

$$= 4.64$$

$$S_{12} = S_{21} = \frac{1}{5}[(-3.2)(-1.6) + (-2.2)(1.4) + (0.2)(-2.6) + (1.8)(3.4) + (3.8)(0.6)]$$

$$= \frac{1}{5}[5.12 - 3.08 + 0.52 + 6.12 - 2.28]$$

$$= 1.28$$

$$\therefore S = \begin{bmatrix} 6.56 & 1.28 \\ 1.28 & 4.64 \end{bmatrix}$$

Now, we assume λ to be eigen value & υ to be eigen vector

$$\Rightarrow S\upsilon = \lambda\upsilon \Rightarrow (S - \lambda)\upsilon = 0$$

$$\Rightarrow \left[\begin{bmatrix} 6.56 & 1.28 \\ 1.28 & 4.64 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right]\upsilon = 0$$

$$\Rightarrow \begin{bmatrix} 6.56 - \lambda & 1.28 \\ 1.28 & 4.64 - \lambda \end{bmatrix} \cdot \upsilon = 0$$

$$\Rightarrow (6.56 - \lambda)(4.64 - \lambda) - (1.28)(1.28) = 0$$

$$\Rightarrow 30.4384 + \lambda^2 - 11.2\lambda - 1.6384 = 0$$

$$\Rightarrow \lambda^2 - 11.2\lambda + 28.8 = 0$$

$$\Rightarrow \boxed{\lambda = 4 \text{ & } \lambda = 7.2} \leftarrow \text{Eigen Values}$$

Now, we calculate Eigen Vectors

$$\begin{bmatrix} 6.56 & 1.28 \\ 1.28 & 4.64 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 7.2 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$6.56x + 1.28y = 7.2x$$
$$1.28x + 4.64y = 7.2y$$

⇒ We can solve any 1 eqⁿ to get values eigen vectors

$$(7.2-6.56)x = 1.28y \Rightarrow \boxed{x = 2y}$$

$$\begin{bmatrix} 6.56 & 1.28 \\ 1.28 & 4.64 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 4 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$6.56x + 1.28y = 4x$$
$$1.28x + 4.64y = 4y$$

$$2.56x = -1.28y \Rightarrow \boxed{-2x = y}$$

∴ Eigen vector $v_1 = \begin{bmatrix} +2 \\ -1 \end{bmatrix}$; $v_2 = \begin{bmatrix} -1 \\ +2 \end{bmatrix}$

$$\Rightarrow V = \begin{bmatrix} +2 & -1 \\ +1 & +2 \end{bmatrix} \Rightarrow V = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$$

We can also transform v_1 & v_2 to unit vectors

$$\Rightarrow v_1 = \begin{bmatrix} 0.89 \\ 0.44 \end{bmatrix} \quad v_2 = \begin{bmatrix} -0.44 \\ 0.89 \end{bmatrix}$$

$$\Rightarrow V = \begin{bmatrix} 0.89 & -0.44 \\ 0.44 & 0.89 \end{bmatrix}$$

Now, we transform data to new co-ordinate space

2	-3.2	-1.6		-3.55 -0.01
3	-2.2	1.4	. [0.89 -0.44 / 0.44 0.89] =	-1.34 2.21
	-0.2	-2.6		-1.32 -2.22
	1.8	3.4		3.09 2.23
	3.8	-0.6		3.11 -2.2

These are the new data values for PC1 & PC2:

Also, if we want to reduce the dimension of dataset then we can take the first column of dataset which will then become 1-D representation of 2-D dataset given to us:

Y : -3.55, -1.34, -1.32, 3.09, 3.11