CS 484 - Introduction to Machine Learning

Assignment - 3

Problem 1:-

A dataset with seven data points is given $\{x_1 \ldots x_7\}$ and the distance between all the pairs are given in a table.

It is also given that clusters $k = 2$

initial clusters centers are : x_3, x_6

so, $c_1 = x_3$ and $c_2 = x_6$

1) **First Iteration :-**

Now, we will be calculating between each points to both x_3 and x_6 and then we will be assigning them to the nearest cluster center

	x_3	x_6	Assigned cluster
x_1	3	2	c_2
x_2	4	7	c_1
x_3	0	5	c_1
x_4	4	1	c_2
x_5	3	8	c_1
x_6	5	0	c_2
x_7	6	1	c_2

After the first iteration we have the below cluster Assignments :

$c_1 : \{x_2, x_3, x_5\}$, $c_2 : \{x_1, x_4, x_6, x_7\}$

2) <u>Second Iteration</u> :-

Here, we should calculate the cluster centers again to get the updated cluster centers

<u>Cluster 1</u> :-

Now, we will calculate the total distance for each data points in this cluster.

For x_2 : $(x_2, x_3) + (x_2, x_5) = 4+1 = 5$

For x_3 : $(x_3, x_2) + (x_3, x_5) = 4+3 = 7$

For x_5 : $(x_5, x_2) + (x_5, x_3) = 1+3 = 4$

So, here x_5 has the lowest distance.
So, It will be the new cluster center.

<u>cluster 2</u> :-

Now, we will calculate the total distance for each data points in this cluster.

For x_1 : $(x_1, x_4) + (x_1, x_6) + (x_1, x_7) = 1+2+3$
$= 6$

For x_4 : $(x_4, x_1) + (x_4, x_6) + (x_4, x_7) = 1+1+2$
$= 4$

For x_6 : $(x_6, x_1) + (x_6, x_4) + (x_6, x_7) = 2+1+1$
$= 4$

For x_7 : $(x_7, x_1) + (x_7, x_4) + (x_7, x_6) = 3+2+1$
$= 6$

Here, two points x_4, x_6 has the lowest total distance. So, we can choose any one of the both. we are choosing x_6 as the new cluster center.

	x_5	x_6	Assigned cluster
x_1	6	2	C_2
x_2	1	7	C_1
x_3	3	5	C_1
x_4	7	1	C_2
x_5	0	8	C_1
x_6	8	0	C_2
x_7	9	1	C_2

After the second iteration, we have the below cluster assignments:

$C_1 : \{x_2, x_3, x_5\}$, $C_2 : \{x_1, x_4, x_6, x_7\}$

3) The algorithm converges when no change occurs in the assignments of clusters.

The two clusters formed when the Lloyd's algorithm converges are:

$C_1 : \{x_2, x_3, x_5\}$

$C_2 : \{x_1, x_4, x_6, x_7\}$

Problem 2 :-

Given that,

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

where, $\sum_{k=1}^{K} \pi_k = 1$; $z = \{z_1, z_2, \dots, z_K\}$

z_k satisfies $z_k \in \{0, 1\}$ & $\sum_{k=1}^{K} z_k = 1$

Also given, the conditional distribution $p(x|z)$ for the observed variable x is given by

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$$

Now, we need to prove that $p(x)$, obtaining by summing $p(z)p(x|z)$ over all possible values of z, is a G.MM.

Step 1:-

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$$

where $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$

Now, we will be expressing $p(z)p(x|z)$

$$p(z)\,p(x|z) = \left(\prod_{k=1}^{K} \pi_k^{z_k}\right) * \left(\prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}\right)$$

$$= \prod_{k=1}^{K} (\pi_k N(x|\mu_k, \Sigma_k))^{z_k}$$

Step 2:- Now, we will be summing over all possible values of z

$$p(x) = \sum_k p(z)\,p(x|z)$$

Here, as $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$ there are K possible configurations of z, where z_k is 1 and the rest are 0.

For each configuration:-

when $z_k = 1$:

$$(\pi_k N(x|\mu_k, \Sigma_k))^1 = \pi_k N(x|\mu_k, \Sigma_k)$$

when $z_k = 0$:

$$(\pi_k N(x|\mu_k, \Sigma_k))^0 = 1$$

Hence, summing over all configurations we get,

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

Hence, it is proved that p(x), obtained by summing p(z)p(x|z) over all possible values of z is a GMM.

Problem 3& 4:-

Problem 3& 4 code, Plots are attached as a separate Pdf.