Homework - 7

Q1 Part 1

1. For each point, calculate the distance to both centroids C_1 and C_2. Assign the point to the cluster of the nearest centroid.

(0,0)

Distance to C_1 :- $d((0,0),(5,1)) = \sqrt{(5-0)^2+(1-0)^2} = \sqrt{26}$

Distance to C_2 :- 6

Nearest centroid : C_1 (since $\sqrt{26} < 6$)

(4,0)

Distance to C_1 :- $\sqrt{2}$

Distance to C_2 :- 2

Nearest centroid : C_1 (since $\sqrt{2} < 2$)

(5,1)

Distance to C_1 :- 0

Distance to C_2 :- $\sqrt{2}$

Nearest centroid : C_1 (since $0 < \sqrt{2}$)

(6,0)

Distance to C_1 :- $\sqrt{2}$

Distance to C_2 :- 0

Nearest centroid : C_2 (since $0 < \sqrt{2}$)

The initial clustering computed by the k-means algorithm is :-

cluster 1 : {1,2,3}

cluster 2 : {4}.

2. ~~initial~~ compute new centroids

For each cluster the new centroid as the mean of all points in the cluster.

The new centroid C_1 is = (3, 1/3)

C_2 is = (6,0)

By reassigning points to the nearest centroids

⇒ cluster 1 (C_1 = (3, 1/3)) ⇒ {1,2}

cluster 2 (C_2 = (6,0)) ⇒ {3,4}

3. After the second iteration
New $C_1 = (2, 0)$
$C_2 = (5.5, 0.5)$
Reassigning points to the new centroids
Cluster 1 = {1}
Cluster 2 = {2, 3, 4}.

4. After the third iteration
New $C_1 = (0, 0)$
$C_2 = (5, 4/3)$
Reassigning points to the new centroids
Cluster 1 = {1}
Cluster 2 = {2, 3, 4}.

5. After the fourth iteration
New $C_1 = (0, 0)$
$C_2 = (5, 4/3)$
Reassigning points to the new centroids
Cluster 1 = {1}
Cluster 2 = {2, 3, 4}.

6. After the fifth iteration
New $C_1 = (0, 0)$
$C_2 = (5, 4/3)$
Since the clustering stabilized after the third iteration and remained the same in the fourth so it will be same in the fifth iteration.
Cluster 1 = {1}
Cluster 2 = {2, 3, 4}.

Q2 Part 1

1. Initial centroids
$C_1 = (1, 0)$, $C_2 = (4, 2)$
Assign points to clusters
P_1, P_2, P_3 are closer to $C_1 = (1, 0)$
P_4, P_5, P_6 are closer to $C_2 = (4, 2)$

So, C(1) = 1, C(2) = 1, C(3) = 1, C(4) = 2, C(5) = 2, C(6) = 2

New centroid $C_1 = (1,1)$

$C_2 = (4,1)$

$$E = \sum_{i=1}^{m} d(P_i, C(P_i))^2$$

$E = 0+1+1+0+1+1 = 4$

$\therefore E = 4.$

2. Initial centroids

$C_1 = (4,1)$, $C_2 = (4,0)$

Assign points to clusters

P_1, P_2, P_4, P_5 are closer to $C_1 = (4,1)$

P_3, P_6 are closer to $C_2 = (4,0)$

Clustering after the first iteration:-

C(1) = 1, C(2) = 1, C(3) = 2, C(4) = 1, C(5) = 1, C(6) = 2.

New centroids $C_1 = (2.5, 1.5)$

$C_2 = (2.5, 0)$

clustering remains the same

$$E = \sum_{i=1}^{m} d(P_i, C(P_i))^2 \Rightarrow 10+5 = 15.$$

PART-2

1. $(1,1) \Rightarrow$ squared distance : $(1-1)^2 + (1-1)^2 = 0.$

$(1,2) \Rightarrow (1-1)^2 + (2-1)^2 = 1.$

$(1,0) \Rightarrow (1-1)^2 + (0-1)^2 = 1$

$(4,1) \Rightarrow (4-1)^2 + (1-1)^2 = 9$

$(4,2) \Rightarrow (4-1)^2 + (2-1)^2 = 10$

$(4,0) \Rightarrow (4-1)^2 + (0-1)^2 = 10$

So, 1d = 0, 2d = 1, 3d = 1, 4d = 9, 5d = 10, 6d = 10.

2. Total sum of squared distances :-

$0+1+1+9+10+10 = 31.$

$(1,1) \Rightarrow P_1 = 0/31 = 0$

$(1,2) \Rightarrow P_2 = 1/31 = 0.0323$

$(1,0) \Rightarrow P_3 = 1/31 = 0.0323$

$(4,1) \Rightarrow P_4 = 9/31 = 0.2903$

$(4,2) \Rightarrow P_5 = 10/31 = 0.3226$

(4,0) ⇒ $P_6 = 10/31 = 0.3226$

So, $P_1 = 0$, $P_2 = 0.3226$, $P_3 = 0.3226$, $P_4 = 0.29032$, $P_5 = 0.32258$, $P_6 = 0.32258$.

3. Initial centroids

$C_1 = (1,1)$, $C_2 = (4,2)$

Clustering after initial assignment

Cluster C_1 = (1,1), (1,2), (1,0)

Cluster C_2 = (4,1), (4,2), (4,0)

Recalculate centroids

$C_1 = (1,1)$ $C_2 = (4,1)$

Reassign points, now the final clustering

C_1 = (1,1), (1,2), (1,0)

C_2 = (4,1), (4,2), (4,0)

So, C(1) = 1, C(2) = 1, C(3) = 1, C(4) = 2, C(5) = 2, C(6) = 2 & E = 4.

4. If the algorithm selects the point with the second-largest probability, it will still choose (4,2) because both P(5) and P(6) are equal.

The clustering process does not change, as the second cluster center remains the same as in the orginal which is (4,2).

Since the cluster centers and assignments remain the same, the quantization error also remains the same.

∴ The answer will not change if the algorithm selects the point with the second largest probability.

5. Selecting the point with the third-largest probability as the second centroid does not change the final clustering or the quantization error. The clusters and quantization error remain the same.

6. To find, we need to sum the probabilities of the three points.

$P_1 = 0$, $P_2 = 0.0323$, $P_3 = 0.0323$, $P_4 = 0.2903$, $P_5 = 0.3226$, $P_6 = 0.3226$

The three points with the largest probabilities are points 4, 5 & 6.

$\therefore \Rightarrow 0.2903 + 0.3226 + 0.3226$

$\Rightarrow 0.9355$

Q3. Count of each salary

Low = 3, Medium = 6, High = 2. & Total instances = 11

P(Low) = 3/11, P(Medium) = 6/11, P(High) = 2/11

P(systems | Low) = 0, P(systems | Medium) = 1/3, P(systems | High) = 1.

After calculating all like this :-

Posterior for Low = 0.

Posterior for Medium = 0.0455

Posterior for High = 0.

∴ The Naive Bayesian classification for the salary of the sample with the values "systems", "senior" and "21-30" is Medium.