Q.2 Consider foll. dataset with min. support count = 60%. Find all frequent itemset using apriori algo & also generate strong association rules if min confidence = 50%.

Transaction ID	Item Bought
T1	{M, O, N, K, E, Y}
T2	{D, O, N, K, E, Y}
T3	{M, A, K, E}
T4	{M, U, C, K, Y}
T5	{C, O, O, K, I, E}

Hint: O bought 4 times in total, but it occurs in just 3 transactions.

Min Support count = min. support threshold * total no. of trans

$= \frac{60}{100} \times 5$

$= 3$

Min Support = 60%.

Database.

TID	ITEM.
T1	MONKEY
T2	DONKEY
T3	MAKE
T4	MUCKY
T5	COOKIE

Scan D →

C1

Itemset	Sup.
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

L_1 →

itemset	Sup
M	3
O	3
K	5
E	4
Y	3

C_2 →

itemset
{M, O}
{M, K}
{M, E}
{M, Y}
{O, K}
{O, E}
{O, Y}
{K, E}
{K, Y}
{E, Y}

C2

itemset	Sup
{M, O}	1
{M, K}	3
{M, E}	2
{M, Y}	2
{O, K}	3
{O, E}	3
{O, Y}	2
{K, E}	4
{K, Y}	3
{E, Y}	2

Scan D →

L_2

itemset	Sup
{M, K}	3
{O, K}	3
{O, E}	3
{K, E}	4
{K, Y}	3

L_3

itemset	~~Sup~~
{O, K, E}	
{K, E, Y}	

Scan D →

itemset	Sup
{O, K, E}	3
{K, E, Y}	2

Frequent items => O, K, E

L = {O, K, E}.

S = {O}, {K} {E} {O, K} {O, E} {K, E}.

S → (L-S)	Support	Confidence	Confidence (%)
{O} → {K, E}	3	3/3	100
{K} → {O, E}	3	3/5	60
{E} → {O, K}	3	3/4	75
{O, K} → {E}	3	3/3	100
{O, E} → {K}	3	3/3	100
{K, E} → {O}	3	3/4	75

∴ All the ass. rules are more than 50%.

∴ Strong association rules are

{O} → {K, E}

{O, K} → {E}

{O, E} → {K}.

{K} → {O, E}

{E} → {O, K}

{K, E} → {O}.

MODULE 4

CLASSIFICATION TECHNIQUES.

(1) Decision Tree Classification.

Ex Create classification model using decision tree:

SR.NO	INCOME	AGE	OWN HOUSE
1	Very high	Young	Yes
2	High	Medium	Yes
3	Low	Young	Rented
4	High	Medium	Yes
5	Very high	Medium	Yes
6	Medium	Young	Yes
7	High	Old	Yes
8	Medium	Medium	Rented
9	Low	Medium	Rented
10	Low	Old	Rented
11	High	Young	Yes
12	Medium	Old	Rented

Soln:- No. of records = 12,
yes = 7, rented = 5

Entropy (own house) = $\sum_{i=1}^{c} -p_i \log_2 p_i$

$= E([7/12], [5/12])$

$= E(0.58, 0.41)$

$= -(0.58 \log_2 0.58) - (0.41 \log_2 0.41)$

$= 0.98$

decrease in entropy after dataset

$G(T,X) = E(T) - E(T,X)$

target class ↓ ↓ Income.

Step 1 for E(Income) we have $E(T,X) = \sum_{c \in X} P(c)E(c)$.

INCOME	YES	RENTED	TOTAL
very high	2	0	2
High	4	0	4
Low	0	3	3
Medium	1	2	3
			12

E (Own house, Income) = p(vh) * E(vh) +
p(h) * E(h) +
p(L) * E(L) +
p(M) * E(m).

$$= \left[\left(\frac{2}{12}\right) \times E\left(\frac{2}{2}, \frac{0}{2}\right)\right] + \left[\left(\frac{4}{12}\right) \times E\left(\frac{4}{4}, \frac{0}{4}\right)\right] + \left[\left(\frac{3}{12}\right) \times E\left(\frac{0}{3}, \frac{3}{3}\right)\right] + \left[\left(\frac{3}{12}\right) \times E\left(\frac{1}{3}, \frac{2}{3}\right)\right].$$

$$= 0 + 0 + 0 + \left[\left(\frac{8}{12}\right) E\left(\frac{1}{3}, \frac{2}{3}\right)\right].$$

$$= 0.25 * \left[(0.33 \log_2 0.33) - (0.67 \log_2 0.67\right.$$

$$= 0.25 * 0.92$$

$$= 0.23$$

Step 2: For E(Age) we have $E(T,x) = \sum_{c \in X} P(c)\, E(c)$

AGE	YES	RENTED	TOTAL
Young	3	1	4
Medium	3	2	5
Old	1	2	3
			12

E(Own house, Age)

$$= p(y) * E(y) + p(m) * E(m) + p(o) * E(o)$$

$$= \left[\left(\frac{4}{12}\right) * E\left(\frac{3}{4}, \frac{1}{4}\right)\right] + \left[\left(\frac{5}{12}\right) * E\left(\frac{3}{5}, \frac{2}{5}\right)\right]$$

$$+ \left[\left(\frac{3}{12}\right) * E\left(\frac{1}{3}, \frac{2}{3}\right)\right]$$

$$= [0.333 * [-(0.75 \log_2 0.75) - (0.25 \log_2 0.25)$$

$$+ [0.416 * [-(0.6 \log_2 0.6) - (0.4 \log_2 0.4)]$$

$$+ [0.25 * [-(0.33 \log_2 0.33) - (0.66 \log_2 0.6)$$

$$= 0.333 * 0.3$$

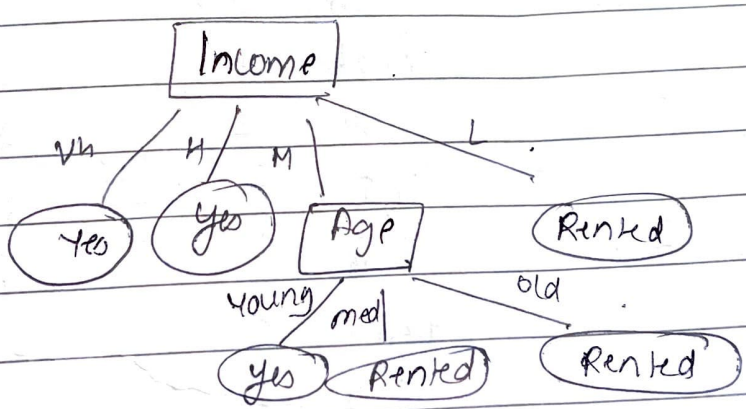
$$= \underline{0.9}$$

Gain (O, I) = E(O) - E(O, I)
= 0.98 - 0.23
= 0.75

Gain (T, X) = E(T) - E(T, X)

Gain (O, A) = E(O) - E(O, A)
= 0.98 - 0.9
= 0.08

Income attribute has highest gain, so used as a decision attribute in the root node.



18_ITB. Tanisha Shaikh

Subject :- DMBI Experiment / Tutorial / Assignment No. :- 05 Page :- 01

Explain K- means clustering & solve the following with K = 3 {2, 3, 6, 8, 9, 12, 15, 18, 22}

Solution :-

K-means clustering is an unsupervised learning algorithm.

K-means algorithm helps to groups the unlabeled dataset into different clusters.

Here; K defines as the number of predefined clusters that need to be created in the process.

It is a centroid based algorithm, where each cluster is associated with a centroid.

The aim of the algorithm is to minimize the sum of distances between the data point and clustering their corresponding cluster.

The algorithm takes the unlabeled dataset as input, divides the dataset into K number of clusters, & repeats the process until it does not find the best clusters.

Given :-

Dataset : {2, 3, 6, 8, 9, 12, 15, 18, 22}

K = 3

Taking the three mean;

$M_1 = 2$; $M_2 = 3$; $M_3 = 6$

Dividing the dataset as per the mean.

$K_1 = \{2\}$

$K_2 = \{3\}$

$K_3 = \{6, 8, 9, 12, 15, 18, 22\}$

Now calculating new mean as per the clusters created.

$M_1 = 2$; $M_2 = 3$; $M_3 = \left(\frac{6+8+9+12+15+18+22}{7}\right) = 12.85$

Dividing the dataset as per the newly created mean;

$K_1 = \{2\}$

$K_2 = \{3, 6, 8, 9\}$

$K_3 = \{12, 15, 18, 22\}$

Now, we will calculate a new mean as per new clusters:

$M_1 = 2$; $M_2 = \left(\frac{3+6+8+9}{4}\right) = 6.5$ $M_3 = \frac{12+15+18+22}{4}$

Dividing the dataset as per the newly created mean;

$K_1 = \{2, 3\}$

$K_2 = \{6, 8, 9\}$

$K_3 = \{12, 15, 18, 22\}$.

Now, we will calculate a new mean as new cluster.

$M_1 = 2.5$; $M_2 = \frac{6+8+9}{3} = 7.67$, $M_3 = \left(\frac{12+15+18+22}{4}\right) =$

$K_1 = \{2, 3\}$

$K_2 = \{6, 8, 9, 12\}$

$K_3 = \{15, 18, 22\}$.

Now, we will calculate the new mean again and form the new clusters

$M_1 = 2.5$ $M_2 = 8.75$ $M_3 = \frac{15 + 18 + 22}{3} = 18.3$

$K_1 = \{2, 3\}$
$K_2 = \{6, 8, 9, 12\}$
$K_3 = \{15, 18, 22\}$.

∴ Finally we got the final 3 clusters as from the given dataset :-

$K_1 = \{2, 3\}$
$K_2 = \{6, 8, 9, 12\}$
$K_3 = \{15, 18, 22\}$.

Group the following objects into $K = 2$ group of medicine based on the two features (ph & weight index).

Obj	Attribute 1 (x)	Attribute 2 (y)
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Let's take the two centroid C1 and C2 as (1, 1) & (2, 1) respectively. The given dataset need to be cluster into 2 groups

Subject :- Experiment / Tutorial / Assignment No. :- Page :-

By Applying the Euclidean distance:

$$D(i,j) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

D^0 =	A	B	C	D	Cluster Centroid
	0	1	3.6	5	C_1 (1,1) - GP1
	1	0	2.82	4.24	C2 (2,1) - GP1

For C_1 (1,1)

$A = \sqrt{(1-1)^2 + (1-1)^2} \Rightarrow 0$

$B = \sqrt{(1-2)^2 + (1-1)^2} \Rightarrow 1$

$C = \sqrt{(1-4)^2 + (0-3)^2} = \sqrt{9+4} = \sqrt{13} = 3.6$

$D = \sqrt{(1-5)^2 + (1-4)^2} = \sqrt{16+9} = \sqrt{25} = 5$

For C(2,1)

$A = \sqrt{(2-1)^2 + (1-1)^2} \Rightarrow 1$

$B = \sqrt{(2-2)^2 + (1-1)^2} \Rightarrow 0$

$C = \sqrt{(2-4)^2 + (1-3)^2} \Rightarrow \sqrt{4+4} \Rightarrow \sqrt{8} \Rightarrow 2.82$

$D = \sqrt{(2-5)^2 + (1-4)^2} = \sqrt{9+9} \Rightarrow \sqrt{18} \Rightarrow 4.24$

The new cluster becomes after finding the distance

$K_1 = \{A\}$

$K_2 = \{B, C, D\}$

Now, the new centroid as per the new clus are:

C_1 (1,1)

$C_2 \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = C_2 (3.67, 2.67)$

Calculating the euclidean distance as per the new centroid :-

D^0 = A	B	C	D	Cluster Centroid
0	1	3.6	5	$C_1(1,1)$
3.15	1.36	0.47	1.88	$C_2(3.67, 2.67)$

The new cluster becomes after finding the distance :

$K_1 = \{A, B\}$

$K_2 = \{C, D\}$

Now, the new centroid as per the new cluster are :

$$C_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = C_1(1.5, 1)$$

$$C_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) \Rightarrow C_2(4.5, 3.5)$$

Calculating the euclidean distance as per the new centroids :

$C_1(1.5, 1)$ & $C_2(4.5, 3.5)$

D^0	A	B	C	D	Cluster Centroid
	0.5	0.5	3.2	4.6	$C_1(1.5, 1)$
	4.3	3.5	0.7	0.7	$C_2(4.5, 3.5)$

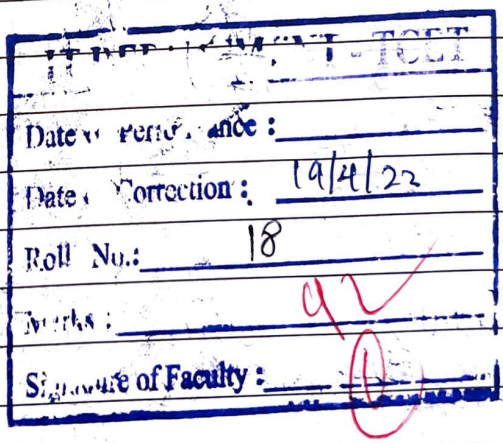
Subject :- Experiment / Tutorial / Assignment No. :- Page :-

The new cluster becomes after finding the distance:

$K_1 = \{A, B\}$
$K_2 = \{C, D\}$

Here we can observe the same cluster got resulted as per the previous iteration

∴ The final clusters groups of medicines are
$K_1 = \{A, B\}$ & $K_2 = \{C, D\}$



18_ ITB_ Tanisha

Subject :- DMBI

Experiment / Tutorial / Assignment No. :- 04

Page :- 01

Unknown sample.

X' = (age = '<=30' ; Income = "median", student = "yes", credit rating = "Fair").

age	Income	Student	credit_rating	buys_computer
<=30	High	no	Fair	no
<=30	high	no	Excellent	no
31..40	high	no	Fair	yes
>40	medium	no	Fair	yes
>40	Low	yes	Fair	yes
>40	Low	yes	excellent	no
31...40	Low	yes	Excellent	yes
<=30	Medium	no	Fair	no
<=30	Low	yes	Fair	yes
>40	Medium	yes	Fair	yes
<=30	Medium	yes	Excellent	yes
31...40	Medium	no	Excellent	yes
31..40	High	yes	Fair	yes
>40	Medium	no	Excellent	no

STEP1 → LEARNING PHASE.

(Buys_computer = Yes) = 9/14

(Buys_computer = No) = 5/14.

age	(buys_computer = Yes)	buys_computer = No
<=30	2/9	3/5
31..40	4/9	0/5
>40	3/9	2/5.

Income	Yes	No
Higher	2/9	2/5
medium	4/9	2/5
Low	3/9	1/5.

Student	Yes	No
no	3/9	4/5
yes	6/9	1/5.

credit_rating	Yes	No
excellent	3/9	3/5
Fair	6/9	2/5.

* Given a new instance

x' (age = '<= 30', Income = "medium", student = "Y
credit rating = "Fair").

Map Rule.

P(x'/Yes) = P(age = '<=30'/Yes) * P(Income = medium
P(student = "yes"/Yes) * P(credit rating

P(X'/No)

= P(age = "<=30"/No) * P(Income = medium/No)
* P(student = Yes/No) * P(credit_rating = fair/No)
* P(No)

$= \left(\frac{3}{5}\right)\left(\frac{2}{5}\right)\left(\frac{1}{5}\right)\left(\frac{2}{5}\right)\left(\frac{5}{14}\right)$

= 0.0068

≈ 0.007

Hence,
we got

P(X'/Yes) = 0.0282.
P(X'/No) = 0.007

Since; 0.028 > 0.007
∴ The Naive Bayesian classifier predicts buyes computer = "yes" for Sample X'.