K.G.C.E.
Karjat - Raigad

Assignment - 1

Page No. :
Date :

KGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCEKGCE

Name: Mohit Sunil Surve

Subject: DMBI

Roll No: 74

Class: TE|IT

D.O.P	D.O.A	Sign	Marks

Q1 A database has ten transactions. Let minimum ~~set ex~~ support = 30% and minimum confidence = 70%. i) Find all frequent patterns using Apriori algorithm
(ii) List strong association rules

Transaction id	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, B, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Soln S1: Generate single item list set:

Support = $\frac{30}{100} \times 10 = 3$

Items	Support
A	6
B	7
C	6
D	7
E	6
F	3
G	2
H	3
I	1
J	4
K	4
L	

Item set	above 30% support
A	6
B	7
C	6
D	7
E	6
F	3
H	3
K	4
L	4

S2: Generate 2 item set :

Item	Support
AB	4
AC	4
AD	4
AE	3
AF	1
AH	2
AK	2
AL	2
BC	5
BD	6
BE	4
BF	1
BH	0
BK	3
BL	2
CD	5
CE	2
CF	1
CH	1
CK	2
CL	1
DE	4
DF	2
DK	1
DL	2
EF	2
EH	3
EK	3
EL	3
FH	3
FK	2
FL	0
HK	2
HL	1
KL	2
	1

Item set above 30% support

~~AB~~	4
AC	4
~~AD~~	4
~~AE~~	3
BC	5
BD	6
~~BE~~	4
~~BK~~	3
CD	5
DE	4
EK	3
~~EK~~	3

S3: Generate 3 item set

Item set of 3 times :-

Item set	Support
ABC	3
ABD	4
ABE	2
ABK	1
ACD	3
ACE	1
ADE	2
AEK	1
AEL	1
BCD	5
BCE	2
BCK	1
BDK	3
BDK	2
BEK	2
BEL	1
CDE	2
DEK	2
DEL	1

Item set above 30% support

Item set	Support
ABC	3
ABD	4
ACD	3
BCD	3
BDE	3

S4: Generate 4 item set

Item set	Support
ABCD	3
ABDE	2
BCDE	2

So, ABCD is the large item set with minimum support 30%. Following rules are generated :

Rule	Confidence	Confidence %
A → ABCD	3/6 = 0.5	50%

B → ACD	3/7 = 0.43	43%
C → ABD	3/6 = 0.5	50%
D → ABC	3/7 = 0.4	43%
AB → CD	3/4 = 0.75	75%
BC → AD	3/5 = 0.6	60%
CD → AB	3/5 = 0.6	60%
AC → BD	3/4 = 0.75	75%
AD → BC	3/4 = 0.75	75%
BCD → A	3/5 = 0.6	60%
ACD → B	3/3 = 1	100%
ABD → C	3/4 = 0.75	75%
ABC → D	3/3 = 1	100%

From above rules generated, only the rules having greater than 70% are considered as final rules so, final rules are

AB → CD
AC → BD
AD → BC
ACD → B
ABD → C
ABC → D

Q2. Use Apriori Algorithm with mini. support count = 2, confidence = 60% to find all frequent itemsets and strong association rules. consider the transaction database given below

Transaction Id	Items
10	1,3,4
20	2,3,5
30	1,2,3,5
40	2,5
50	1,3,5

Soln S1: Generate single item set

Item set	Support
1	3
2	3
3	4
4	1
5	4

Item set above 2 support count

1	3
2	3
3	4
5	4

S2: Generate 2 time set

Item set	Support
1,2	1
1,3	3
1,4	1
1,5	2
2,3	2
2,5	3
3,4	1
3,5	3

item-set above 2 support count

item set	Support

1,3	3
1,5	2
2,3	2
2,5	3
3,5	3

S3: Generate 3-item set

item-set	support
1,2,3	1
1,2,5	1
2,3,5	2
1,3,4	1
1,3,5	2

{1,3,5}, {2,3,5} is large item-set with minimum support 12'.

Following rules are generated:

Rule	Confidence	Confidence %
2^3 = 5	2/2 = 1	100%
3^5 = 2	2/3 = 0.66	66%
2^5 = 3	2/3 = 0.66	66%
2^3^5	2/3 = 0.66	66%
3 = 2^5	2/4 = 0.5	50%
3 = 2^3	2/4 = 0.5	50%

from above generated rules, only rules having 60% are considered as final rules.

∴ final rules are

2^3 = 5

3 ^ 5 = 2

2 ^ 5 = 3

2 = 3 ^ 5

Rules generated

Rules	Confidence	Confidence %
1 ^ 3 = 5	2/3 = 0.66	66%
3 ^ 5 = 1	2/3 = 0.66	66%
1 ^ 5 = 3	2/2 = 1	100%
1 = 3 ^ 5	2/3 = 0.66	66%
3 = 1 ^ 5	2/4 = 0.5	50%
5 = 1 ^ 3	2/4 = 0.5	50%

∴ final rules 1 ^ 3 = 5

3 ^ 5 = 1

1 ^ 5 = 3

1 = 3 ^ 5

Q3 Apply Naive Bayes classfier algo, to classify an unknown sample X (outlook = Sunny, Temp = Cool, Humidity = High, windy = false).

Sample dataset is given as follows:-

Outlook	Temp	Humidity	windy	Class
Sunny	Hot	high	false	N
Sunny	Hot	high	true	N
Overcast	Hot	high	false	P
Rain	mild	high	false	P
Rain	cool	normal	false	P
Rain	Cool	normal	true	N

Overcast	cool	normal	true	P
Sunny	mild	high	true	N
Sunny	cool	normal	false	P
Rain	mild	normal	false	P
Sunny	mild	normal	false	P
Overcast	mild	high	true	P
Overcast	hot	normal	True	P
Rain	mild	high	false	N

Soln Given a training set we can compute the probabilities as follows:

- The classification prim may be formalized using a posterior posibilities : P(C|T) is probability of the ~~pos~~ sample tuple $Y = \langle y_1, y_2, \ldots, y_n \rangle$ is of C.

P(C|Y)) is maximal

from the above given sample data, calculate probability for play tennis (P) and tennis (N) for all attribute.

Outlook

P(sunny|yes) = 2/9	P(sunny|NO) = 3/5	
P(overcast|yes) = 4/9	(overcas|NO) = 0	P(yes) = 9/14
P(rain|yes = 3/9	P(rain|NO) = 2/5	P(NO) = 5/14

Temp

P(hot|yes) = 2/9	P(hot|NO) = 2/5
P(mild|yes) = 4/9	P(mild|NO) = 2/5
P(cold|yes) = 3/9	P(cold|NO) = 1/5

Humidity

P(high|yes) = 3/9	P(high|NO) = 4/5
P(False|yes) = 6/9	P(False|NO) = 2/5

An unseen sample Y = (sunny, hot, high, false)

P(Y|yes, P(yes) = P(rain|yes) · P(hot|yes) · P(high|yes)
· P(false|yes) · P(NO)
= 3/9 · 2/9 · 3/9 · 6/9 · 9/14
= 0.010582

Choose the class so that it maximises the probability. This means new instance can be classified as no sample Y is classified as class No.

An unsee sample = <sunny, cool, high, true>

P(Y|yes) · P(yes) = P(sunny|yes) · P(cool|yes) · P(high|yes) · P(true|yes) · P(yes)
= 2/9 · 3/9 · 3/9 · 9/14
= 0.0053.

P(NO) · P(Y/NO) = P(sunny|NO) · P(cool|NO) · P(high|NO) · P(true/No) · P(NO)
= 3/5 · 1/5 · 4/5 · 3/5 · 5/14
= 0.0206

Now, choose the class so that it maximize this probability. This means that the new instance will be classified as no.

Q4. Define classification discuss the issues in classification. A simple ex. from stock market involving only discrete ranges has profit as categorical attribute, with value {up down} and the data is

Age	competition	Type	Profit
old	yes	software	Down
old	no	software	Down
old	no	software	Down

mid	yes	Software	Down
mid	yes	hardware	Down
mid	No	hardware	Down
mid	No	Software	Up
new	yes	software	Up
new	No	hardware	Up
new	No	Software	Up

Soln Classification:

- It is the form of data analysis
- classification constructs classification model based on training data set.
- Using this model it classifies the new data.
- classification model it classifies the predicat.
- categorical class labels
- for e.g we can classify model to categories bank loan application as either safe or risky.

Issues in classification:

① Data cleaning - removing the noise.
② ~~Reve~~ Relevance analysis - Also & relevant attributes
③ Data transformation & reduction.
  a) Normalization - Involves scaling all values.
  b) Generalization - transformed by generation.

class = Profit p=up , n=down

Total no. of records = 10

∴ p = 5 & n = 5

So info. gain = $I(p,n) = \frac{-p}{p+n} \log_2 \frac{p}{p+n} = \frac{n}{p+n} \log_2$

$$\therefore I(5,5) = \frac{-5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

$$I(5,5) = 1$$

S1. Compute entropy for age : cold, mild, new) for age = old
p_i with 'up' class = 0 & n_i with 'down' class = 3

$$I(p_i, n_i) = I(0,3) = \frac{-0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\frac{3}{3} = 0$$

Similarly for different age ranges $I(p_i, n_i)$ are calculated as follows:

Age	p_i	n_i	$I(p_i, n_i)$
old	0	3	0
mid	2	2	1
new	3	0	0

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(Age) = \frac{3}{10}(0) + \frac{4}{10}(1) + \frac{3}{10}(0) = 0.4$$

$$Gain(I, age) = I(p,n) - E(age)$$
$$= 1 - 0.4 = 0.6$$

Just as use calculated entropy & gain for age, same can be calculated for competition & type. as follows

Entropy can be competition (yes, NO)

competition ~~can~~ be compe'	p_i	n_i	$I(p_i, n_i)$
yes	1	3	0.811
NO	4	2	0.918

$E(\text{competition}) = \frac{4}{10} \times (0.811) + \frac{6}{10}(0.918) = 0.875$

Gain (T, competition) = $1 - 0.875$

$= 0.125$

Entropy of type : (S|w, H|w):

Type	p_i	n_i	$I(p_i, n_i)$
S|w	3	3	1
H|w	2	2	1

$E(\text{type}) = \frac{6}{10}(1) + \frac{4}{10} \times 1 = 1$

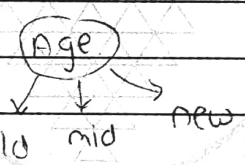
$\therefore G(T, \text{type}) = 1 - 1 = 0$

Gain (T, age) = 0.6

Gain (T, competition) = 0.125

Gain (T, type) = 0.

$\therefore$ we get



S2: Since attribute age as not, we have to decide remaining two attributes for age brance node.

Age	Contest	Type	Profit
old	yes	S|w	down
old	NO	S|w	down
old	No	H|w	down

$\therefore$ The profit is down for every old age is simple branched as down.

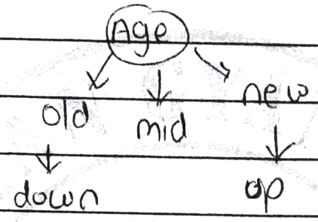
S3: consider age = New

Age	competition	Type	Profit

New	yes	S/W	up
New	NO	H/W	up
New	No	S/W	up

∴ profit is up for every new age is simply branched as up.

So, new decision tree becomes



S4: Now consider age = mid.

Age	competition	Type	profit
mid	yes	S/W	Down
mid	yes	H/W	Down
mid	No	H/W	up
mid	NO	S/W	up

(i) compute Entropy for competition : (yes, No)

$p = 1$, $n = 2$

competition	p_i	n_i	$I(p_i, n_i)$
yes	0	2	0
NO	2	0	0

∴ Entropy (competition) = $\frac{2}{4}(0) + \frac{2}{4}(0) = 0$

Gain (age, competition) = 1 - 0 = 1

(ii) compute Entropy for type (S/W, H/W)

Type	pi	ni	I(pi, ni)
S|w	1	1	1
H|w	1	1	1

$$E(type) = \frac{2}{4} \times (1) \quad T \quad \frac{2}{4} \times (1)$$

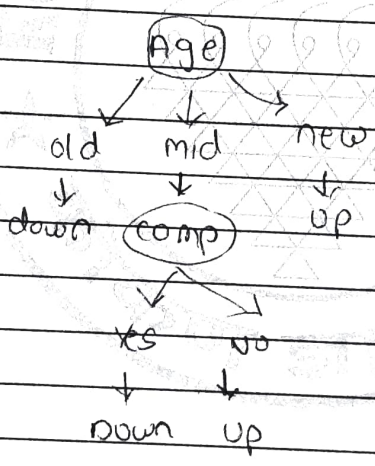
$$= 1$$

$\therefore$ G(Tage, Type) = 1 - 1 = 0

Gain (Tage, compt^n) = 1

Gain (Tage, Type) = 0

So, we got decision tree as follows



Q5 What is clustering? Explain K-means clustering algorithm, suppose data for clustering is {2, 4, 10, 12, 03, 20, 11, 5, 25, 36, 41, 14}. Assuming number of clusters to be 2 i.e k=2, cluster given data using above algorithm.

Soln - Clustering is an unsupervised learning problem.

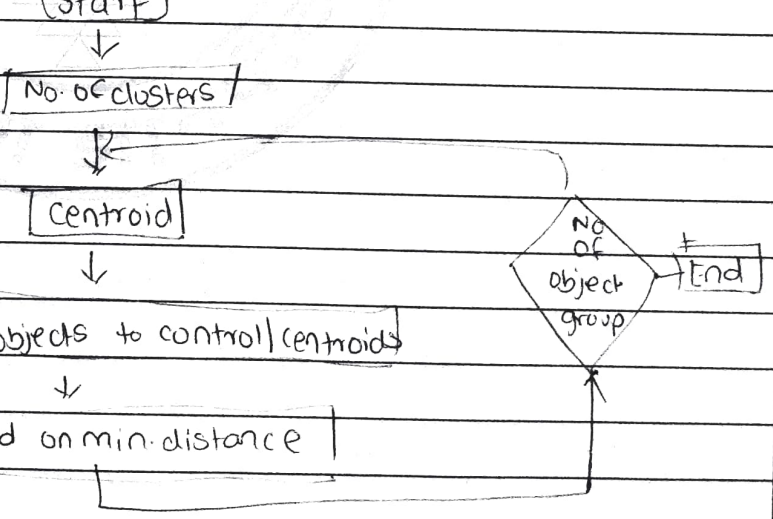
- Clustering is a data mining technique used to place data.

elements into related groups without advanced knowledge of group definition.

- Its a process of partitioning data objects into sub classes which are called as clusters
- A cluster contains data objects which have as high inter similarity and low intra similarity.

K-means clustering algorithm:

- K-means clustering algorithm is one of the partitioning method. Its simplest unsupervised learning methods.
- K-means clustering aims to partition n' observations into k' clusters in which each observation belongs to the clusters with the nearest mean, serving as a prototype of the cluster.
- This results in partitioning of data space.
- K is +ve integer number.



Flowchart of K-means

Above figure shows flowchart of K-means clustering

- Define K centroids for K clusters which are generally far only from each other.
- Then group elements into clusters which are nearer to the centroid of that cluster.
- After this 1st step, calculate new centroid from each cluster based on elements of that cluster.
- Follow the same method & group the elements based on new centroid.

Given :- {2, 4, 10, 12, 3, 20, 11, 25, 5, 36, 41, 14} & K = 2.

Randomly assign means : $m_1 = 4$, $m_2 = 12$

Again calculate new mean for new cluster groups.

(1) $K_1 = \{2, 4, 3, 5\}$ $K_2 = \{10, 12, 20, 11, 25, 36, 41, 14\}$
$m_1 = 5$ $m_2 = 21$

(2) $K_1 = \{2, 4, 3, 5, 10, 12, 11\}$ $K_2 = \{20, 25, 36, 41, 14\}$
$m_1 = 7$ $m_2 = 27$

(3) $K_1 = \{2, 4, 3, 5, 10, 12, 11, 14\}$ $K_3 = \{20, 25, 36, 41\}$
$m_1 = 8$ $m_2 = 30$

(4) $K_1 = \{2, 4, 3, 5, 10, 12, 11, 14\}$ $K_2 = \{20, 25, 36, 41\}$

Stop as clusters in step 3 & 4 are same. clusters in last two groups are identical.

So, $K_1 = \{2, 4, 3, 5, 10, 12, 11, 14\}$ and $K_2 = \{20, 25, 36, 41\}$

Q6. Explain Business Intelligence Issues.

→ Organizations and people :

1. Management within an organization do not agree that the decision taken based on data or evidence work for them, they prefer to run the operation from instinct.

2. In order to access business progress there is no overall business strategy laid out with objectives and measures for those objective.
3. The data needed for business intelligence system, cannot be obtained from IT personal as they accomplished and they have no resources available.
4. The performance at the business either mating ace of BI or not ther is no incentive provided to the staff within the organization.
5. There is no obvious time to establish a BI system. The busines is is in a static at high chance or flux.

2) Data and Technology:

1. The data at the organization is not clean the time and effort needed to correct or handle this type at data leads to an unsuccessful delivery at the BI support. For e.g. there could be many different destinations for the same item.

e.g. a customer may be coded differently or sales system to that held in accounts system.

2. The technology chosen for BI turns out to be so particular that it ultimately routs in time consuming process which leads to delay in project completion.
3. The BI technology discourages the use of system due to the following reasons:

- The info. presentation quality is poor or limited.
- The response time for data presentation is slow.
- It is too difficult for data presentation is to cut new question of BI teonology for either of them and users of BI expect