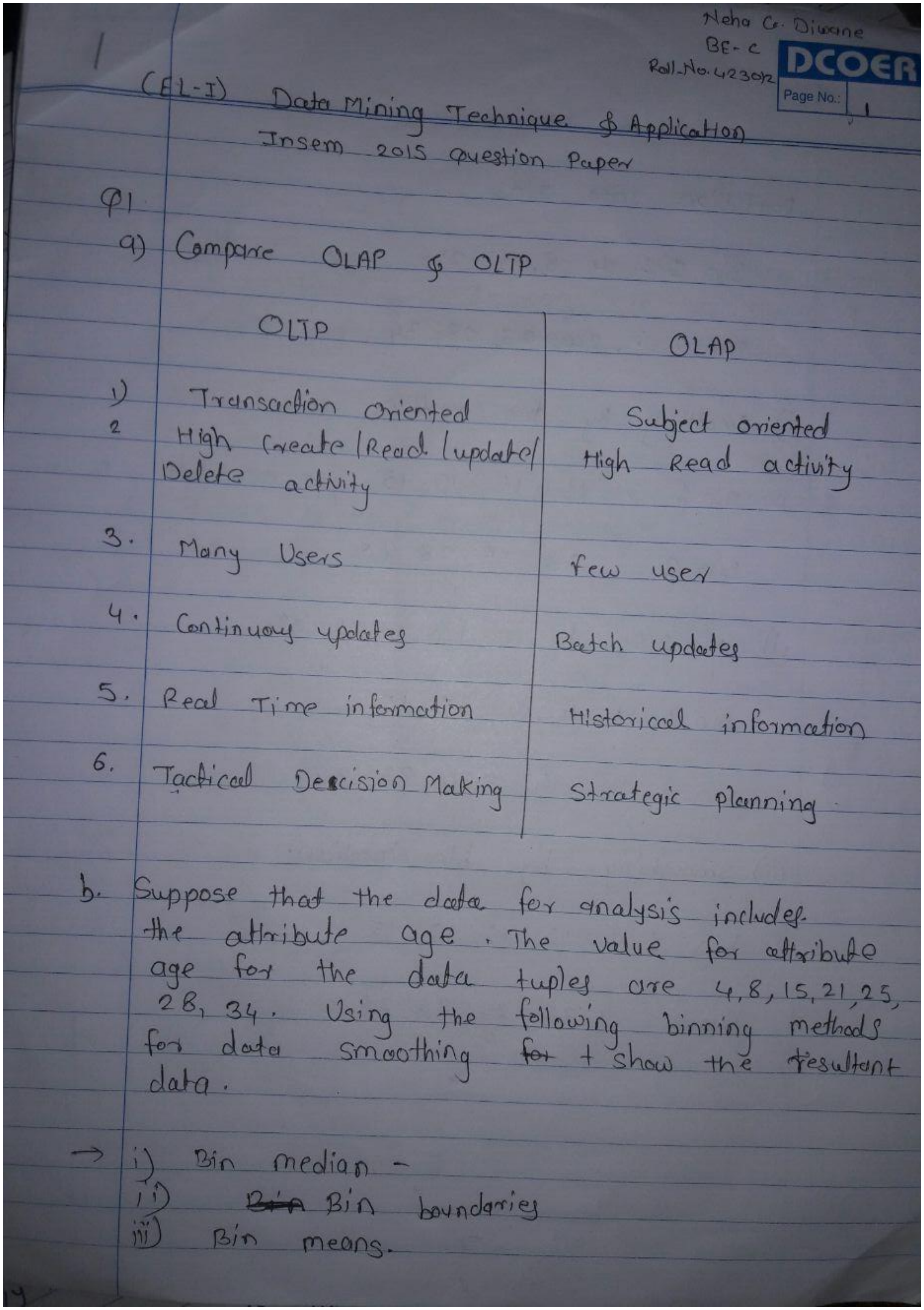
DMTA Solve QP



Neha G. Diwane
BE-C
Roll-No. 423012

DCOER
Page No.: 1

(EL-I) Data Mining Technique & Application
Insem 2015 Question Paper

Q1

a) Compare OLAP & OLTP

	OLTP	OLAP
1)	Transaction Oriented	Subject oriented
2	High Create/Read/update/Delete activity	High Read activity
3.	Many Users	few user
4.	Continuous updates	Batch updates
5.	Real Time information	Historical information
6.	Tactical Decision Making	Strategic planning

b. Suppose that the data for analysis includes the attribute age. The value for attribute age for the data tuples are 4, 8, 15, 21, 25, 28, 34. Using the following binning methods for data smoothing ~~for~~ + show the resultant data.

→ i) Bin median -
ii) ~~Bin~~ Bin boundaries
iii) Bin means.

Partition into Bins

Bin 1 : 4, 8, 15, 21

Bin 2 : 24, 25, 28, 34

i) smoothing by bin median

Bin 1 : 15, 15, 15, 15

Bin 2 : 28, 28, 28, 28.

ii) Smoothing by ~~min~~ bin boundaries.

Bin 1 : 4, 4, 15, 15

Bin 2 : 24, 24, 24, 34

iii) Smoothing by bin median

bin 1 : 15, 15, 15, 15

bin 2 : 25, 25, 25, 25

Q 2

a With suitable ~~eo~~ diagram explain various steps of a knowledge discovery in db process & breifly explain each step.

→ KDD Process -

The process of discovering knowledge in data & application of data mining methods refers to the term knowledge Discovery in databases.

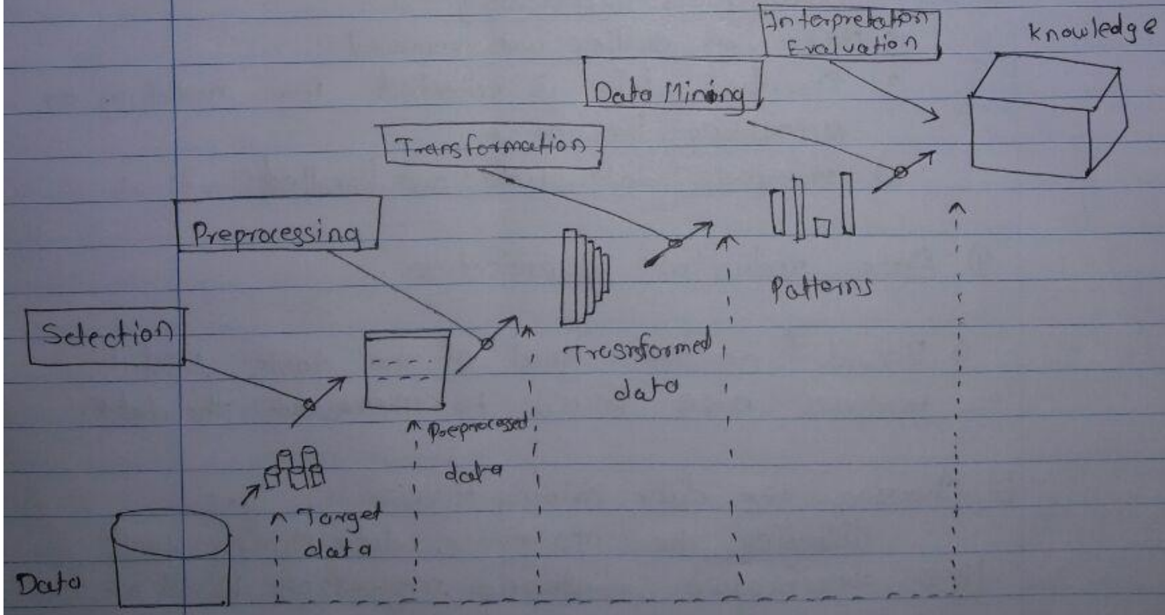


Fig. KDD process.

1) Developing an understanding of :.

a) The application domain
b) The relevant prior knowledge
c) The goals of the end user

2) Creating a target data set. -
Selecting a data set, or focusing on subset of variable or data samples on which discovery is to be performed.

3) Data cleaning & preprocessing -
1) Noise or outlier are removed
2) Essential info. is collected for modeling or accounting for noise
3) missind data field are handled.

4) Data reduction & projection.

1) Based on the goal of the task, Useful features are found to represent the data.

5) choosing the data mining task -
Selecting the appropriate data Mining task like classification, clustering, regression based on the goal of the KDP process.

b. What is predictive & discriptive Data Mining.

→ ~~predictive data Mining~~ -

Discriptive data mining. -

- The data range of continuous attribute is divided into intervals.
- Categorical attribute are accepted by only afew classification algorithm.
- By discritization the size of the data is reduced & prepared for further analysis.

c. Explain ~~in~~ brief

Q3

a) W.r.t. Association Rule Mining Define.

i) Support -

The support of an itemset is the count of that itemset in the total no. of transaction or in ~~*~~ other words it is the percentage of the transaction in which the item appear

if $A \Rightarrow B$

$$\text{Suppor}(A \Rightarrow B) = \frac{\#\text{-tuple Containing both } A \,\&\, B}{\text{total } \sout{\text{of}} \text{ no. of tuples.}}$$

ii) Confidence :.

The Confidence ~~or length~~ strength for an association rule $A \Rightarrow B$ is the ratio of the no. of transactions that $A \cup B$ to the no. of transactions that Contain A.

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{tuples Containing both } A \,\&\, B}{\text{tuple Containing } A.}$$

b A db has five transactions : let min sup. = 60%. & min conf = 80%.

TID	Item-Bought
T100	{MONKEY}
T200	{DONKEY}
T300	{MAKE}
T400	{MUCKY}
T500	{COOKIE}

Find all frequent itemset using Apriori Algo.

C_1

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

ii min support count = 2

MO	1	M	= 3
ME	2	O	- 3
MK	2	E	= 4
MY	2	K	= 4
OE	3	Y	= 3
OK	2		
OY	2		
EK	3		
EY	2		
KY	2		

$L_2 = OE = 3$

$EK = 3$

iii) $C_3 = L_2 \bowtie L_2$

$OEK = 2$

iii) $O \rightarrow E = \frac{O \cup E}{O} = 3/5$

$E \rightarrow O = 3/5$

$E \rightarrow K = 4/5 = 75\%$

$K \rightarrow E = 4/5 = 75\%$

Q4

a) what do you mean by basket analysis?
How it can help a grocery shopper?

→ Market basket analysis is a modeling technique which is also called as affinity analysis it helps identifying which items to be purchased to other.

- Market Basket analysis problem assumes we have some large no. of items eg. "bread", "milk" etc ~~et~~ Customer has taken together So, the marketers use the information to put items.

The problem of large volume of trival result can be overcome with the help of differential market basket of enables in finding interesting result & eliminate large volume

- Some special observation among the rule eg. if the rule which holds in one store but not in any other than it thing may be really interesting to note that there is some-thing special about that store in the way it has organized.

b) Is the support & confidence of an association $X \rightarrow Y$. The Same as the of $Y \rightarrow x$. Why or not if an itemset of 'n' item as frequent are all subset of this frequent itemset necessary frequent ?

→ - An itemset is closed if none of its immediate superset has same support as the itemset.

- Consider two itemset x & y if every time of x in y but there is atleast one item of Y. which is not in X, then y is non proper set of x. incage of itemset X.

- If an itemset x is minimal frequent itemset or max itemset if x is frequent & there exist no super item y such that x is subset of y & y is frequent.

- To find frequence itemsed one can use the monotocity principle.

c> Explain apriori algorithm for generating association rules. What is time complexity.

→ - A Apriori algorithm solves the frequent itemset problem.

- A algorithm analysis a data set to determine which combinations of item occur together frequenty.

- A aproiry algo is the core of the various algo. for data mining problem. The best known problem is finding the association rules that holds in a basket.

Input - D, a db of transactions. min-sup

Output - L, frequent intemset in D

methods -

```
L1 = find frequent itemset.
for ($k=2$ , $Ik = 2$, $Lk = 2$, $L_{k-1} \neq \phi$ , k++)){
$C_k$ = aprori-gen ($L_{k-1}$) ;
for each transaction $t \in D$
$C_t$ = subset (ck-t) ;
for each candidate $c \in C$ ;
c.count ++ ;
}
$L_k = \{ c \in C_k \mid c.count \geq$ min-sup)
}
return $L = \cup_k . L_k$
```

Q5

a) State bayes theorem -

→ Baye's theorem is used find conditional probabilities
- The conditional probability of an event is a likehood obtained with the conditional info that some other event has previously occured.
P(x|y) is conditional probability of an event occuring for the event before Y which has already occured

$$P(x/y) = P(x \& y) / P(A)$$

- An initial probability called a priori probability which we get before additional info is obtained.

b) Apply ID3 on the following training dataset from all electronics customer database & extracting classification.

Age	income	Student	Credit-rating	class-buys-Comp
<=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
30...40	low	yes	excellent	yes.

Age	income	student	Credit-rating	class-bays-Comp
<=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

→ for age <=30

p_i = with "yes" class = 2 & n_i = with "No" class = 3

$\therefore I(p_i, n_i) = I(2,3) = 0.971$

age	p_i	n_i	(p_i, n_i)
<=30	2	3	0.971
31....40	4	0	0
>40	3	2	0.971

entropy from value table,

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

$$E(age) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$$

$$= 0.694$$

Hence, gain (age) = I(P, n) − E(age)

= 0940 − 0.694 = 0.246

gain (income) = 0.029

gain (student) = 0.151

gain (credit-rating) = 0.048

a) Entropy for income,

$I(P_i, n_i) = I(0, 2) = (0/2) \log (0/2) - 2/2 \log 2/2 =$ 0

Income	Pi	ni	I (Pi, ni)
High	0	2	0
medium	1	1	1
low	1	0	0

$E(A) = (2/5) * I(0,2) + (2/5) * I(1,1) + (1/5) * I(1,0)$

$= 0.4$

gain (s<=40, income) = I(P, n) − E(income)

= 0.971 − 0.4

= 0.571

b) Calculate entropy for student = (No. Yes)

for student = No

p_i = with 'yes' class = 0 & n_i = with No class = 3

$I(P_i, n_i) = I(0, 3) = (0/3) \log (0/3) - (3/3) \log (3/3) = 0$

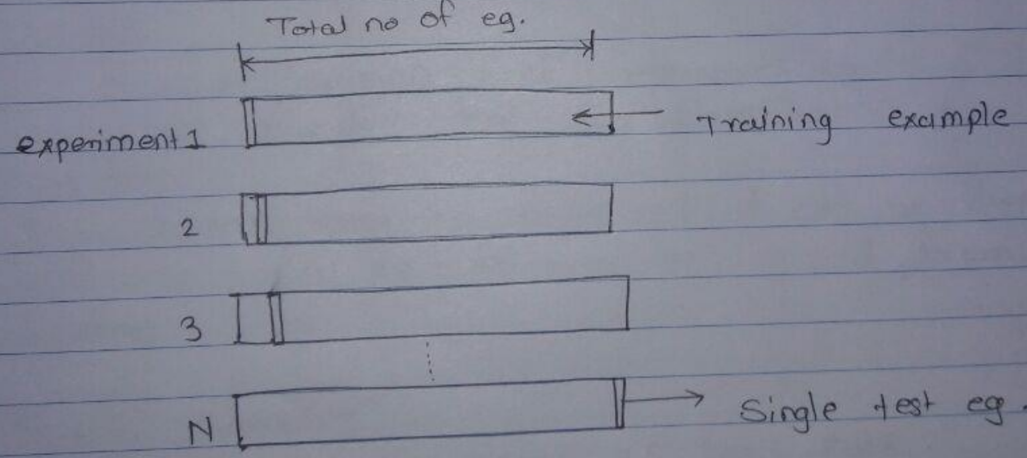
Student	Pi	ni	I (Pi ni)
No	0	3	0
yes	2	0	0

C. Explain the term 10-fold ~~v~~ cross validation. what is the significance of it

→ - This gives accurate estimate of evaluation.

- The estimate's variance get reduced due to stratification.

- Ten-fold cross validation is ~~p~~ repeated ten times & finally the results are averaged based on the previous 10 result.



Neha G. Niwane

Q6

a) Several different classifier such as Bayes, Descision Tree, KNN are available. State various performance metrics that are Used to evaluate the the classifiers. Compare the above three classifiers using metrics

→ Descision Tree -

- Training data set should be class based for learning of descision tree in descision tree induction.
- The ~~a~~ decision tree represents rules & it is very a popular tool for classification & prediction
- Rules are easy to understand & can be directly used in SQL to retrieve to record from database.

Baye's Theorem -

- It is also known as Bay's Rule
- Bay's theorem is used to find Conditional probabilities.

~~event~~ - The Conditional probabilities of an event is a likehood obtained with the additional information that some other event has previously occured.

$$p(X|Y) = P(X \text{ and } Y) / P(A)$$

→ Hair

P(Blonde | yes) = 2/3	P(Blonde No) = 1/5
P(Brown | yes) = 0	P(Bown No) = 4/5
P(Red | yes) = 1/3	P(Red | No) = 0

Height

P(Avg | yes) = 2/3	P(Avg | No) = 0
P(Tall | yes) = 0	P(Tall | No) = 2/5
P(short | yes) = 1/3	P(short | No) = 2/5

Weight

P(righ | yes) = 1/3	P(light | No) = 1/5
P(Avg | yes) = 1/3	P(Avg | No) = 2/5

location

P(No | yes) = 3/3	P(No | No) = 2/5
P(yes | yes) = 0	P(yes | No) = 3/5

P(yes) = 3/8
P(No) = 5/8

An unseen X = < brown , tall , average , no >

P(X | yes) . P(yes) = P(Brown | yes) . P(tall yes)
. P(avg | yes) . P(No | yes)
. P(yes) = 0

P(X | No) . P(No) = P(Brown | No) . P(tall | No)
. P(avg | No) . P(No | No) . P(No)
= 0.032

Since 0.032 > 0 , our eg. gets classified as No