Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

Association Mining

Association rule

- Let us assume that the number of items in the shop stocks is n.
- i.e. In the dataset there 6 items in stock, namely Bread, Milk, Diapers, Beer, eggs, Cola.
- The items list is represented I and its items represented by {i1, i2, ---- in).
- The number of transactions are represented by N transactions N=5 for the shop data given.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5.	{Bread, Milk, Diapers, Cola}.

Each transaction is denoted by T ($t_1, t_2 \cdots t_N$) with a unique identifier (TID) and each transaction consists of a subset of items (Possibly a small subset) purchased by one customer.

- Let each transaction of M items be (i1, i2, ... in), where m <= n
- m <= n (number of items in a transaction should be less than or equal to total number of items in the shop).
- Typically, transactions differ in the number items.

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

From our dataset

T_1 = has two items ($m=2$) with i_1 - bread & i_2 = milk

T_4 = $m=4$, i_1 - Bread, i_2 - Milk, i_3 = Diapers, i_4 = Beer

- Our task will be to find association relationships, given a large number of transactions, such that items that tend to occur together are identified.

Note: Association rules mining does not take into account the quantities of items bought.

- Metrics to Evaluate the strength of Association Rules
  - Support, Confidence, lift.
  - Let N be the total number of transactions
  - Support of X is represented as the number of time X appears in the database divided by N.
  - While support of X and Y together is represented as the number of times they appear together devided by N as given below.

$$\text{Support}(X) = \frac{\text{Number of time } X \text{ appears}}{N} = P(X)$$

$$\text{Support}(XY) = \frac{\text{Number of time } X \text{ and } Y \text{ appears together}}{N}$$

$$= P(X \cap Y).$$

Thus support of X is the probability of X while support of XY is the probability of $X \cap Y$.

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

Support (Bread) = Number of times Bread appears / Total number of transactions.

= 4/5 = P(Bread)

Support (milk) = 4/5 P(milk)

Support (Diapers) = 4/5 P(Diapers)

Support (Beer) = 3/5 P(Beer)

Support (Eggs) = 1/5 P(Eggs)

Support Cola = 2/5 = P(Cola)

Support (Bread, milk) = Number of times Bread, Milk appear together / Total number of transactions

= 3/5 = P(Bread ∩ milk)

Support (Diapers, Beer) = Number of times Diapers, Beer / No of transactions

= 3/5 = P(Diapers ∩ Beer)

- A high level of support indicates that the rule is frequent enough for the business to take interest in it.

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

(2) Confidence.

Suppose the support that support of $X \rightarrow Y$ is 80 %. then

It means that $X \rightarrow Y$ is very frequent and there are 80 % chances that X and Y will appear together in a transaction.

- This would of interest to the sales Manager.
- Suppose another pair of items (A and B) and support of $A \rightarrow B$ is 50 %.
- This is not as frequent as $X \rightarrow Y$ but if this was higher such as whenever A appears then is 90% chances that B also appears, then of course it would be of great interest.
- Thus, not only the probability that A and B appear together matters, but also the conditional probability of B when A has already occured play a significant role.
- This conditional probability that B will follow when A has already occured is considered during determining the confidence of the rule.

* The support and confidence are important metrics to judge the quality of association

- A high level of confidence shows that the rule is true often enough to justify a decision based on it.

- Confidence for $X \rightarrow Y$ is defined as the ratio of the support of X and Y together to the support for X. (It is same as the conditional probability of Y when X has a

occurred.

Remark If X appears much more frequently than Y and appearing together, then the confidence will be low.

Confidence of $(X \rightarrow Y)$ = $\frac{\text{Support}(XY)}{\text{Support}(X}$

$= \frac{P(X \cap Y)}{PX} = P(Y|X)$

~~Confidence~~ $P(Y|X)$ is the probability of Y once X has taken place, also called the conditional probability.

Examples of the support measure.

TID	Items
1	ABC.
2	ABD.
3	BC.
4	AC.
5	BCD.

Support = Occurrence / Total support.

Total support = 5

Support{AB} = 3/5 = 40%

Support{BC} = 3/5 = 60%

Support{ABC} = 1/5 = 20%

1. Given $X \Rightarrow Y$
2. Confidence = Occurrence {X and Y} / Occ(X)
3. $A \Rightarrow B = 2/3$
4. $B \Rightarrow C = 3/4$
5. $AB \Rightarrow C$ 1/2
6.

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

Database for identification of association rules.

Antecedent	Consequent
A	0
A	0
A	1
A	0
B	1
B	0
B	1

We are two rules derived from the association of the combine

Rule 1 : A Implies 0 i.e A → 0

Rule 2 : B Implies 1 i.e B → 1

The support for rule = Number of time A & 0 appear

Total Number of transaction

Support of Rule = 3/7

Support for rule 2 = 2/7

Confidence for Rule 1 is support of (A,0) / Sup

Support of (A,0) = 3/7 = 3/4

Support (A) = 4/7 = 3/7 / 4/7 = 3/7 × 7/4 = 3/4

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

Confidence for Rule 2

= support (B, I) / support (B).

support (B, I) = 2/7 = 2/7

= support of B) = 3/7 = 3/7

2/7 X 7/3 = 2/3.

Lift

- It is very important to consider the frequency of Y or P(Y) for the effectiveness of the association mining rule $X \rightarrow Y$
- Confidence is the conditional probability of Y when X has already occured.
- It is very important to consider how frequent Y is to gauge the effectiveness of the ~~association mining~~ confidence.

Thus

Lift is the ration of conditional probability of Y when X is given to the unconditional probability of Y in the dataset.

- Simply.

It is the confidence of $X \rightarrow Y$ divided by the probability of Y

Registration No. ..... Question .....

Write on both sides of the paper

Do not ... in this mar...

Lift = P(Y/X)/PY

Or

Lift = Confidence of (X → Y) / P(Y)

Or

Lift = (P(X ∩ Y) / P(X)) / P(Y)

Thus lift can be Computed by dividing the Confidence by the unconditional probability of Consequent Y.

- Coke is a very common sales item in a store and appears in most of the transactions.
- Candle → Coke = has support of 20%
  Confidence of 90%.
- logically we think that is very popular and it appears in 95% of transactions.
- Then obviously it also appears quite often with the Candle as well.
- So the rule of association of Candle and Coke will not be useful.

- Candle → Matchbox — support 20%
  — Confidence 90%.

Suppose the frequency of Matchbox is very little Compared to the sale of Coke.

- Rule suggest that when we make a sale of Candles, 90% chances indicate a matchbox will be sold in the same transaction.

Registration No. ..... Question .....

Write on both sides of the paper

Do not write in this margin

It is more effective and logical to conclude that when we sell a candle then we also sell a Coke (Coke is popular and will appear with every item not just with Candle). As support and confidence are unable to handle this case. It is handled by the lift.

In the case Probability of Y is very low (in the case Candle → Matchbox (Y Matchbox)

Probability of Y is very high Candle → Coke (Item Y, is Coke)

Lift = $P(Y/X) / P(Y)$

or

Lift = Confidence of $(X \rightarrow Y) / P(Y)$

or

Lift = $(P(X \cap Y) / P(X)) / P(Y)$

- High probability of Y i.e $P(Y)$ makes lift less effective and to low value makes it more effective

Lift for Rule 1 $A \rightarrow O$

$P(A,O)$ 3/4, $P(A)$ 4/4 $P(O)$ 4/4

Confidence of $A \rightarrow O$ = 3/4 / 4/4 $\quad 3/4 \times 4/4 = \frac{21}{16}$

$= 1.312$

Registration No. ..... Question .....

Do not write in this margin

Write on both sides of the paper

Lift for Rule 2 $B \rightarrow 1$ $= \frac{3/3}{3/4}$ $= \frac{3}{3} \times \frac{4}{3} = \frac{12}{9} = 1.5$

Confidence $(2/3) / P_1 (3/4)$

Confidence for Rule 1 is $3/4 = 0.75$

Rule 2 is $2/3 = 0.66$

It should be Observed that although Rule 1 has a higher Confidence as compared to Rule 2, It has a lower lift as compared to Rule 2.

Naturally,
It would appear that Rule 1 is more valuable because of having higher confidence; it appear more accurate. But the accuracy of the rule can be misleading if it is independent of the dataset.

Lift, as a metric is important because it considers both the confidence of the rule and the overall dataset