Assignment 4: Solution just for reference

1. A Contingency Table for X → Y is defined as follows (where X' signifies Not X, and likewise for Y):

	Y	Y'	
X	f_{11}	f_{10}	f_{1+}
X'	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$\|T\|$

f_{11}: support of X and Y

f_{10}: support of X and Y'

f_{01}: support of X' and Y

f_{00}: support of X' and Y'

Suppose we are given the following contingency table

	Coffee	Coffee'	
Tea	15	5	20
Tea'	75	5	80
	90	10	100

(a) Determine if confidence is a useful measure for the Rule

Tea → Coffee

Now, consider another measure called Lift, defined as follows

$$Lift\ (A, B) = \frac{P(A|B)}{P(A)}$$

(b) Comment on this measure for the cases of

1. Lift = 1,
2. Lift > 1,
3. 0< Lift < 1.
4. Lift = 0.

(c) Calculate the Lift of the contingency table of Tea and Coffee, and comment on its usefulness in relation to confidence.

<u>Soln</u>

(a) Association Rule: Tea → Coffee
Confidence= σ(Tea, Coffee)/ σ(Tea) = 15/20= 0.75

but P(Coffee) = 0.9

and P(Coffee|Tea') = 75/80= 0.9375

Although confidence is high, the rule Tea → Coffee is misleading, since the probability of someone drinking coffee given he/she does not drink tea is 93.75%

Therefore confidence is not a very useful measure.

(b) Note that

$$Lift\ (A, B) = \frac{P(A|B)}{P(A)} = \frac{P(A \cap B)}{P(A)P(B)}$$

- P(A ∩ B) = P(A) × P(B) => Statistical independence [Lift = 1]
- P(A ∩ B) > P(A) × P(B) => Positively correlated [Lift > 1]

- P(A ∩ B) < P(A) × P(B) => Negatively correlated [0<Lift < 1]
- Lift = 0 suggests that Events A and B are mutually exclusive

(c)

$$Lift = \frac{0.15}{(0.2)(0.9)} = 0.833$$

which is less than 1. Therefore tea and coffee are negatively correlated. This is more useful than confidence.

2. Consider the following two contingency tables from Information Retrieval counting the number of documents that contain both words X and Y.

   Is Lift a good measure here?

X=Compiler, Y=Mining

	Y	Y'	
X	10	0	10
X'	0	90	90
	10	90	100

	Y	Y'	
X	90	0	90
X'	0	10	10
	90	10	100

X=Data, Y=Mining

Solution

For first table

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

For second table

$$Interest = \frac{0.9}{(0.9)(0.9)} = 1.11$$

- We expect the words "data" and 'mining" to appear together more frequently than the words "complier" and "mining" in a collection of computer science articles. However, the above results suggest that "compiler" and "mining" appear together more often than the words "complier" and "mining".

  Therefore, lift is not a good measure here.

3. Consider the data set shown in the table below. Suppose we are interested in extracting the following association rule:

$$\{\alpha_1 \leq \text{Age} \leq \alpha_2, \text{Play Piano} = \text{Yes}\} \rightarrow \{\text{Enjoy Classical Music} = \text{Yes}\}$$

Age	Play Piano	Enjoy Classical Music
9	Yes	Yes
11	Yes	Yes
14	Yes	No
17	Yes	No
19	Yes	Yes
21	No	No

25	No	No
29	Yes	Yes
33	No	No
39	No	Yes
41	No	No
47	No	Yes

To handle the numerical attribute, we apply the equal-frequency approach with 3, 4, and 6 intervals. Categorical attributes are handled by introducing as many attributes as the number of categorical values. Assume that the support threshold is 10% and the confidence threshold is 70%.

(a) Suppose we discretize the Age attribute into 3 equal-frequency intervals. Find a pair of values for α_1 and α_2 that satisfy the minimum support and minimum confidence requirements.

(b) Repeat part (a) by discretizing the Age attribute into 4 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

(c) Repeat part (a) by discretizing the Age attribute into 6 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

(d) From the results in part (a), (b), and (c), discuss how the choice of discretization intervals will affect the rules extracted by association rule mining algorithms.

Solution

(a) Suppose we discretize the Age attribute into 3 equal-frequency intervals. Find a pair of values for $\alpha 1$ and $\alpha 2$ that satisfy the minimum support and minimum confidence requirements.

Answer:

($\alpha_1 = 9$, $\alpha_2 = 17$): $s = 2/12 = 16.7\%$, $c = 2/4 = 50\%$.

($\alpha_1 = 19$, $\alpha_2 = 29$): $s = 2/12 = 16.7\%$, $c = 100\%$.

($\alpha_1 = 33$, $\alpha_2 = 47$): $s = 0/12 = 0\%$, $c = 0/0$ (i.e., no information can be obtained)

The pair of values that satisfies the minimum support and minimum confidence requirements is (19, 29).

(c) Repeat part (a) by discretizing the Age attribute into 4 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

Answer:

($\alpha_1 = 9$, $\alpha_2 = 14$): $s = 2/12 = 16.7\%$, $c = 2/3 = 67\%$.

($\alpha_1 = 17$, $\alpha_2 = 21$): $s = 1/12 = 8.3\%$, $c = ½ = 50\%$.

($\alpha_1 = 25$, $\alpha_2 = 33$): $s = 1/12 = 8.3\%$, $c = 1/1 = 100\%$

($\alpha_1 = 39$, $\alpha_2 = 47$): $s = 0/12 = 0\%$, $c = 0/0$ (i.e., no information can be obtained)

No rule satisfies the support and confidence thresholds.

(d) Repeat part (a) by discretizing the Age attribute into 6 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

Answer:

($\alpha_1 = 9$, $\alpha_2 = 11$): $s = 2/12 = 16.7\%$, $c = 2/2 = 100\%$.

($\alpha_1 = 14$, $\alpha_2 = 17$): $s = 0/12 = 0\%$, $c = 0/2 = 0\%$.

($\alpha_1 = 19$, $\alpha_2 = 21$): $s = 1/12 = 8.3\%$, $c = 1/1 = 100\%$

($\alpha_1 = 25$, $\alpha_2 = 29$): $s = 1/12 = 8.3\%$, $c = 1/1 = 100\%$

($\alpha_1 = 33$, $\alpha_2 = 39$): $s = 0/12 = 0\%$, $c = 0/0$ (i.e., no information can be obtained)

($\alpha_1 = 41$, $\alpha_2 = 47$): $s = 0/12 = 0\%$, $c = 0/0$ (i.e., no information can be obtained)

(d) From the results in part (a), (b), and (c), discuss how the choice of discretization intervals will affect the rules extracted by association rule mining algorithms.

If the discretization interval is too wide, some rules may not have enough confidence to be detected by the algorithm. If the discretization interval is too narrow, the rule in part (a) will be lost.

Solution to Question 4

(a)

$5\times(2^4-1) = 75$

(b)

In general, for a database having N items, we have

$$\sum_{k=1}^{N-1}\binom{N}{k}\times\binom{N-k}{1}=\sum_{k=1}^{N-1}\binom{N}{k}\times(N-k)$$

Where the first part within the summand corresponds to the LHS of the rule, and the other part corresponds to the single item on the RHS. This is also equal to $N\times(2^{(N-1)}-1)$, which can be seen as follows. First, fixed one item on the RHS, and that will leave (N–1) items on the LHS. Thus, the total number of choices on the LHS is $(2^{(N-1)}-1)$, and “1” needs to be subtracted because the possibility of having no item included in the LHS is not permitted. Varying this for all N items on the RHS, we have $N\times(2^{(N-1)}-1)$.

Solution to Question 5

Here Ck signifies candidate itemsets of size k, and Lk, frequent itemsets of size k, with

A denoting Apple etc.

(a)

C1	
Item	Support
A	5
B	4
C	5
D	2
E	2

This gives L1 as

L1	
Item	Support
A	5
B	4
C	5

since anything less than 3 will not give 50% support of the 6 transactions.

Likewise, we have

C1	
Item	Support
A	5
B	4
C	5
D	2

E	2

This gives L1 as

L1	
Item	Support
A	5
B	4
C	5

since anything less than 3 will not give 50% support of the 6 transactions.

Likewise, we have

C2	
Item	Support
A B	3
A C	4
B C	3

This gives

L2	
Item	Support
A B	3
A C	4
B C	3

Continuing, we have

C3	
Item	Support

A B C	2

Since this does not have enough support, we terminate the process here.

The answer required is therefore L2, which gives the following rules (having minimum support of 50%):

A → B ,   B → A

A → C ,   C → A

B → C ,   C → B

(b)

The confidence of these are:

A → B   (3/5)   B → A (3/4)

A → C   (4/5)   C → A   (4/5)

B → C   (3/4)   C → B   (3/5)

From these, only the following rules has a confidence > 70%, and these are the only rules that meet the required criteria:

B → A (3/4)

A → C   (4/5)

C → A   (4/5)

B → C   (3/4)