CSC 3170 Assignment 4

This is an individual assignment and should be submitted by 5 pm, 12 May 2023 via Blackboard

Question 1

Consider a table *R* with *k*+1 columns, where the first *k* columns $C_1, \ldots\ C_i, \ldots, C_k$ are categorical (i.e., non-numerical) attributes and the last column is a numerical attribute. Let the number of possible values of C_i be n_i, $i = 1, 2, \ldots, k$. Consider the query

select $C_1, \ldots\ C_i, \ldots, C_k$, sum($C_{k+1}$)
from *R*
group by cube($C_1, \ldots\ C_i, \ldots, C_k$);

Let the table resulting from this query be *S*.

Determine the number of tuples in

(i) the table *R*
(ii) the table *S*

Suppose the above query is replaced by

select $C_1, \ldots\ C_i, \ldots, C_k$, sum($C_{k+1}$)
from *R*
group by rollup($C_1, \ldots\ C_i, \ldots, C_k$);

Let the table resulting from this query be *T*.

(iii) provide a formula that gives the total number of tuples in *T*.

Question 2

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_{o+}
	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

(i) Lift = 1,

(ii) Lift > 1,

(iii) 0 < Lift < 1,

(iv) Lift = 0.

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

Question 3

Consider the following two contingency tables obtained from an Information Retrieval application, which counts the number of documents in a repository that contain both words X and Y.

Discuss whether Lift is 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

Question 4

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

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

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.

Questions 5 and 6 are related to the following supermarket basket analysis situation.

Consider a database of customer transactions where a number of items are purchased. In general, the number of possible association rules in such a database is very large, giving rise to a huge amount of processing in support and confidence evaluations. In more complex associations, rules can be of the form

$$\boldsymbol{A_1} \,\&\, \boldsymbol{A_2} \,\&\, \ldots \,\&\, \boldsymbol{A_n} \rightarrow \boldsymbol{B}, \qquad\qquad (*)$$

where the antecedent can be a conjunction of several items, but the consequent is a single item. Consider a database of customer transactions where a number of items are purchased as shown in the table below, with the first column indicating the Transaction ID, and the second column giving the items purchased.

Transaction#	**Items List**
T100	Apple, Beer, Eggs
T200	Apple, Cake, Diaper
T300	Apple, Cake
T400	Beer, Cake
T500	Apple, Beer, Cake
T600	Apple, Beer, Cake, Diaper, Eggs

Question 5

(a) How many possible rules of the form (*) are there for the above database?

(b) In general, for a database where the item list has a total of *N* items, determine the total number of possible rules of the form (*).

Question 6

(a) For the above database, determine all rules having a minimum support of 50% using the *Apriori Algorithm*, giving the support of each rule.

(b) Suppose further that we are only interested in rules having a confidence of at least 70%. Determine the set of rules having minimum support of 50%, and minimum confidence of 70% for this database.