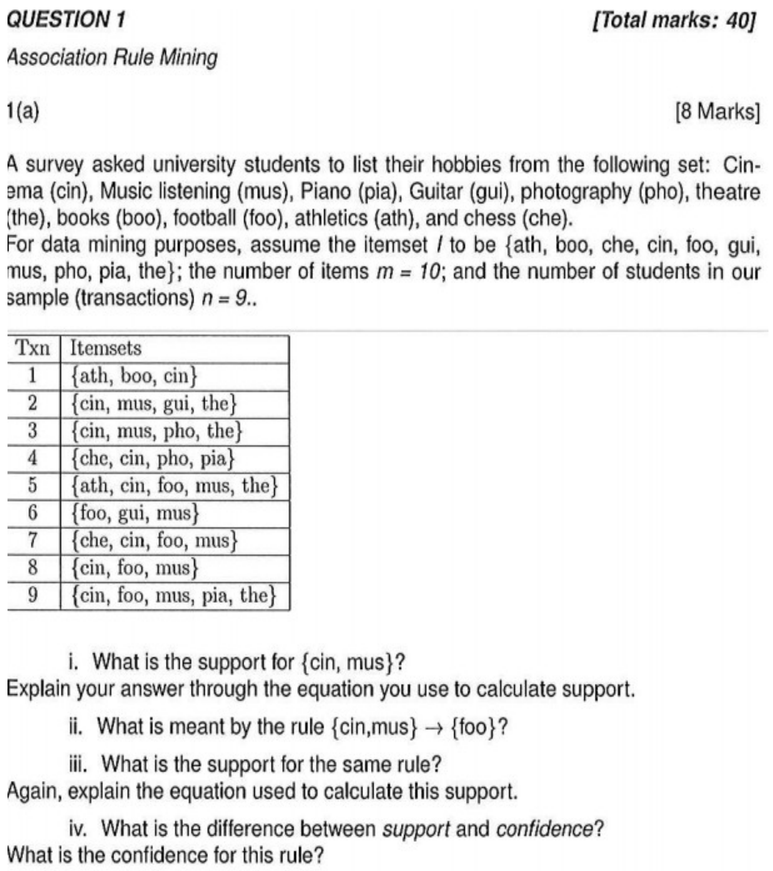
2015 repeat Data Mining



QUESTION 1 *[Total marks: 40]*

Association Rule Mining

1(a) [8 Marks]

A survey asked university students to list their hobbies from the following set: Cinema (cin), Music listening (mus), Piano (pia), Guitar (gui), photography (pho), theatre (the), books (boo), football (foo), athletics (ath), and chess (che).
For data mining purposes, assume the itemset *I* to be {ath, boo, che, cin, foo, gui, mus, pho, pia, the}; the number of items *m = 10*; and the number of students in our sample (transactions) *n = 9*..

Txn	Itemsets
1	{ath, boo, cin}
2	{cin, mus, gui, the}
3	{cin, mus, pho, the}
4	{che, cin, pho, pia}
5	{ath, cin, foo, mus, the}
6	{foo, gui, mus}
7	{che, cin, foo, mus}
8	{cin, foo, mus}
9	{cin, foo, mus, pia, the}

i. What is the support for {cin, mus}?
Explain your answer through the equation you use to calculate support.

ii. What is meant by the rule {cin,mus} → {foo}?

iii. What is the support for the same rule?
Again, explain the equation used to calculate this support.

iv. What is the difference between *support* and *confidence*?
What is the confidence for this rule?

a)
I)
support for an itemset is the proportion of transactions that contain all the items in S defined as:

$Support(S)=Count(S)/n$

Where Count(S) = number of transactions matching S

$Support(\{cin, mus\})=Count(\{cin, mus\})/n=6/9=0.66$

ii)
the rule {cin,mus} → {foo} means that when cin and mus are chosen we can predict foo will also be chosen.

Iii)
Support for a rule L → R is the proportion of transactions in which the items in both L and R occur together defined as:

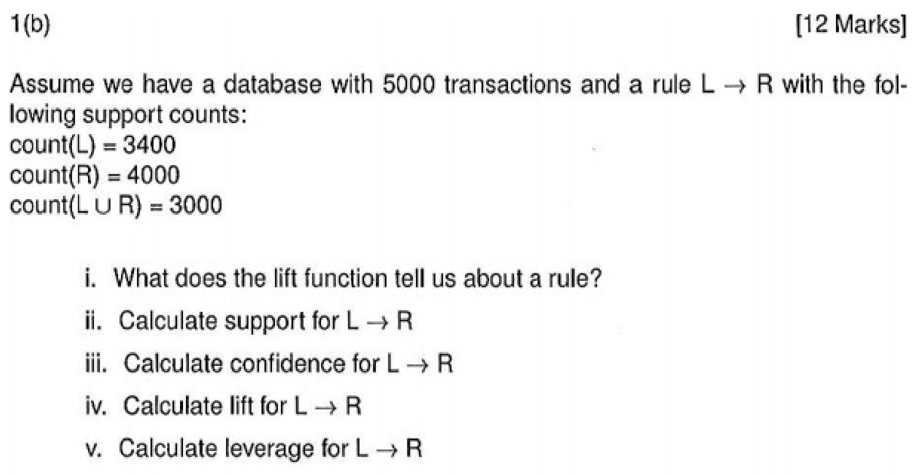
$Support(L \rightarrow R)=Count(L \cup R)/n$

$Support(\{cin, mus\} \rightarrow \{foo\})=Count(\{cin, mus\} \cup \{foo\})/n=4/9=0.44$

iv)
Support applies to all transactions in the database, whereas confidence is the proportion of transactions for which the rule is satisfied and is defined as:

$Confidence(S \rightarrow R)=Count(L \cup R)/Count(L)$

$Confidence(\{cin, mus\} \rightarrow \{foo\})=Count(\{cin, mus, foo\})/Count(\{cin, mus\})=4/6=0.66$



1(b) [12 Marks]

Assume we have a database with 5000 transactions and a rule L → R with the following support counts:
count(L) = 3400
count(R) = 4000
count(L ∪ R) = 3000

i. What does the lift function tell us about a rule?

ii. Calculate support for L → R

iii. Calculate confidence for L → R

iv. Calculate lift for L → R

v. Calculate leverage for L → R

b)

I)

The lift function tells us how interesting a rule may be:

$$lift(L \rightarrow R) = Count \frac{(L \cup R)}{Count(L) x \, Support(R)}$$

ii)

$$Support(L \rightarrow R) = 3000/5000 = 0.6$$

iii)

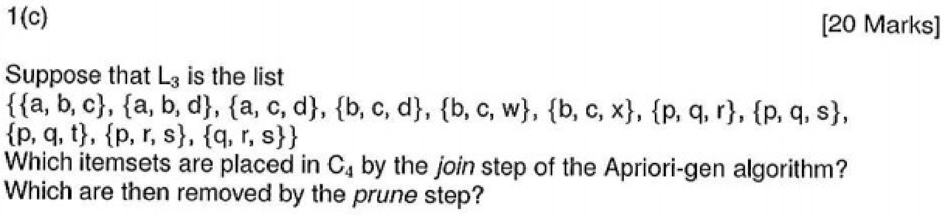
$$Confidence(L \rightarrow R) = 3000/3400 = 0.88$$

iv)

$$Lift(L \rightarrow R) = \frac{3000}{3400 \times (4000/5000)} = 1.10$$

v)

$$Leverage(L \rightarrow R) = Support(L \cup R) - support(L) \times support(R)$$
$$= 0.6 - (3400/5000) \times (4000/5000) = 0.056$$



1(c) [20 Marks]

Suppose that L_3 is the list
{{a, b, c}, {a, b, d}, {a, c, d}, {b, c, d}, {b, c, w}, {b, c, x}, {p, q, r}, {p, q, s}, {p, q, t}, {p, r, s}, {q, r, s}}
Which itemsets are placed in C_4 by the *join* step of the Apriori-gen algorithm?
Which are then removed by the *prune* step?

c)
Join Step
k = 4, k-2 = 2

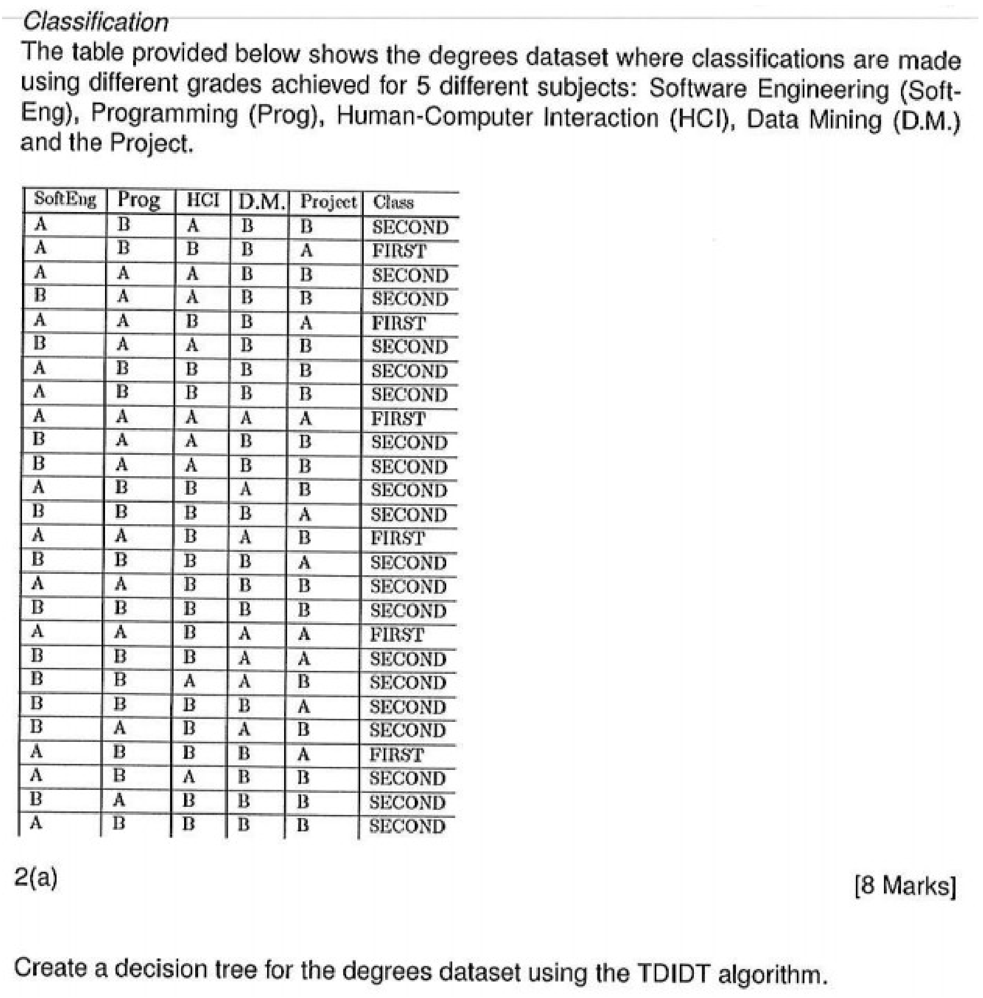
First itemset	Second itemset	Contribution to C
{a,b,c}	{a,b,d}	{a,b,c,d}
{b,c,d}	{b,c,w}	{b,c,d,w}
{b,c,d}	{b,c,x}	{b,c,d,x}
{p,q,r}	{p,q,s}	{p,q,r,s}
{p,q,r}	{p,q,t}	{p,q,r,t}

C_4 = {{a,b,c,d},{b,c,d,w},{b,c,d,x},{p,q,r,s},{p,q,r,t}}

Prune Step

ItemsSet in C_4	subsets	Subsets all in L3
{a,b,c,d}	{a,b,c},{a,c,d},{a,b,d}{b,c,d}	YES
{b,c,d,w}	{b,c,d},{b,d,w},{b,d,w},{c,d,w}	NO
{b,c,d,x}	{b,c,d},{b,c,x},{b,d,x},{c,d,x}	NO
{p,q,r,s}	{p,q,r},{p,r,s},{p,q,s},{q,r,s}	YES
{p,q,r,t}	{p,q,r},{p,r,t}	NO

C_4 = {{a,b,c,d},{p,q,r,s}}



Classification

The table provided below shows the degrees dataset where classifications are made using different grades achieved for 5 different subjects: Software Engineering (Soft-Eng), Programming (Prog), Human-Computer Interaction (HCI), Data Mining (D.M.) and the Project.

SoftEng	Prog	HCI	D.M.	Project	Class
A	B	A	B	B	SECOND
A	B	B	B	A	FIRST
A	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	A	B	B	A	FIRST
B	A	A	B	B	SECOND
A	B	B	B	B	SECOND
A	B	B	B	B	SECOND
A	A	A	A	A	FIRST
B	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	B	B	A	B	SECOND
B	B	B	B	A	SECOND
A	A	B	A	B	FIRST
B	B	B	B	A	SECOND
A	A	B	B	B	SECOND
B	B	B	B	B	SECOND
A	A	B	A	A	FIRST
B	B	B	A	A	SECOND
B	B	A	A	B	SECOND
B	B	B	B	A	SECOND
B	A	B	A	B	SECOND
A	B	B	B	A	FIRST
A	B	A	B	B	SECOND
B	A	B	B	B	SECOND
A	B	B	B	B	SECOND

2(a) [8 Marks]

Create a decision tree for the degrees dataset using the TDIDT algorithm.

Same as 2014 Q4 a