CS422 HW3

1.1 Chapter 5 Exercises: 2,6,8,9,12,13,20

2. a. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

Support Count for {e} = # of transactions containing e = 8
Support for {e} = Support count/ total transactions = (8/10) = **0.8**

Support Count for {b,d} = # of transactions containing b & d = 2
Support for {b,d} = = (2/10) = **0.2**

Support Count for {b,d,e} = # of transactions containing b, d, & e = 2
Support for {b,d,e} = (2/10) = **0.2**

b. Use the results in part (a) to compute the confidence for the association rules {b, d}→{e} and {e}→{b, d}. Is confidence a symmetric measure?
Confidence for {b,d} --> {e} = Support for {b,d,e} / Support for {b,d} = (0.2/0.2) = **1**
Confidence for {e} --> {b,d} = Support for {b,d,e} / Support for {e} = (0.2/0.8)= **0.25**
No, **confidence is not a symmetric measure**.

c. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise).

Support Count for {e} = # of transactions containing e = 4
Support for {e} = Support count/ total transactions = (4/5) = **0.8**

Support Count for {b,d} = # of transactions containing b & d = 5
Support for {b,d} = (5/5) = **1**

Support Count for {b,d,e} = # of transactions containing b, d, & e = 4
Support for {b,d,e} = (4/5) = **0.8**

d. Use the results in part (c) to compute the confidence for the association rules {b, d}→{e} and {e}→{b, d}.
Confidence for {b,d} --> {e} = Support for {b,d,e} / Support for {b,d} = (0.8/1) = **0.8**
Confidence for {e} --> {b,d} = Support for {b,d,e} / Support for {e} = (0.8/0.8) = **1**

e. Suppose s1 and c1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s2 and c2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s1 and s2 or c1 and c2.
There are no apparent relationships between s1, s2, c1, and c2.

6.

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?
There are six items in the data set. Therefore the total number of rules Total # of association rules = 3^d - 2^{d+1} + 1 = 3^6 - 2^7 + 1 = 602 rules.

(b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?
Because the longest transaction contains 4 items, the maxi- mum size of frequent itemset is 4

(c) Write an expression for the maximum number of size-3 itemsets that

can be derived from this data set.
(6,3) = 20.
(d) Find an itemset (of size 2 or larger) that has the largest support.
{Bread, Butter}.
(e) Find a pair of items, a and b, such that the rules {a} −→ {b} and {b} −→ {a} have the same confidence.

Rules	Confidence
Milk→beer	1/5
Beer→milk	1/4
Milk→diapers	4/5
Diapers→milk	4/7
Milk→bread	3/5
Bread→milk	3/5
Milk→butter	2/5
Butter→milk	2/5
Milk→cookies	1/5
Cookies→milk	1/4
Beer→diapers	¾
Diapers→beer	3/7
Beer/bread	0
Bread/beer	0
Beer→butter	0
Butter→beer	0
Beer→cookies	2/4
Cookies→beer	2/4
Diapers→bread	2/7
Bread→diapers	2/5
Diapers→butter	3/7
Butter→diapers	3/5
Diapers→cookies	1/7
Cookies→diapers	¼
Bread→butter	1
Butter→bread	1
Bread→cookies	1/5
Cookies→bread	¼
Butter→cookies	1/5
Cookies→butter	1/4

Here Item sets {Milk, Bread}, {Milk, Butter}, {Beer, Cookies} and {Bread, Butter} have the same confidence. Item sets {Beer, Bread} and {Beer, Butter} also have same confidence, but its 0.
8 Consider the following set of frequent 3-itemsets:
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the data set.

a. List all candidate 4-itemsets obtained by a candidate generation procedure using the Fk−1×F1 merging strategy.

{1,2,3} + {4} = {1,2,3,4}
{1,2,3} + {5} = {1,2,3,5}
{1,2,4} + {5} = {1,2,4,5}
{1,3,4} + {5} = {1,3,4,5}
{2,3,4} + {5} = {2,3,4,5}

Therefore, {1, 2, 3, 4},{1, 2, 3, 5}, {1, 2, 4, 5},{1, 3, 4, 5},{2, 3, 4, 5}.

b. List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.
{1,2,3} vs {1,2,4} - The first 2 position are the same, thus we merge the two.
{1,2,3,4}
{1,2,3} vs {1,2,5} - The first 2 positions are the same, thus we merge. {1,2,3,5}
{1,2,3} vs {1,3,4} - The first 2 positions are not the same, so no merge done.
{1,2,3} vs {1,3,5} - No merge. Because repeat {1,2,3,5}
{1,2,3} vs {2,3,4} - No merge. Because repeat {1,2,3,4}
{1,2,4} vs {1,2,5} - Merge. {1,2,4,5}.
{2,3,4} vs {2,3,5} - Merge. {2,3,4,5}

Therefore, {1,2,3,4}, {1,2,3,5}, {1,2,4,5} {2,3,4,5}.

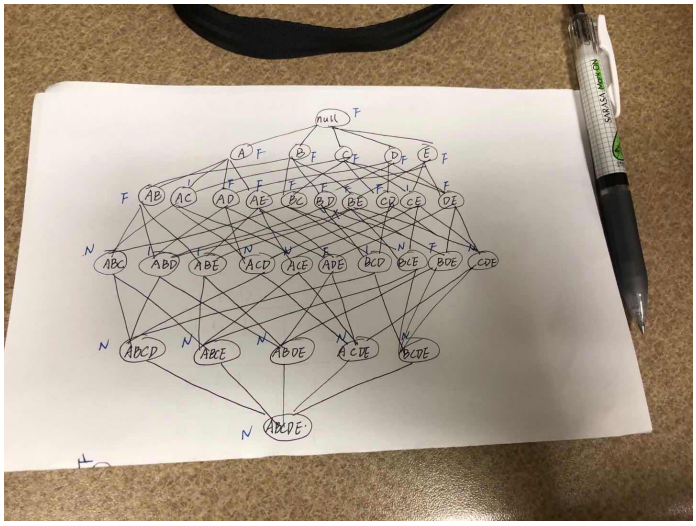
c. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.
{1,2,3,4} - Check if {1,2,3} is a frequent itemset --> Yes.
Check if {1,2,4} is a frequent itemset --> Yes.
Check if {1,3,4} is a frequent itemset --> Yes.
Check if {2,3,4} is a frequent itemset --> Yes.
So this 4-itemset is fine.
{1,2,3,5} - Check {1,2,3} --> Yes.
Check {1,2,5} --> Yes.
Check {1,3,5} --> Yes.
Check {2,3,5} --> Yes.
So this 4-itemset is fine.
{1,2,4,5} - Check {1,2,4} --> Yes.
Check {1,2,5} --> Yes.
Check {1,4,5} --> No.
This 4-itemset is out.
{2,3,4,5} - Check {2,3,4} --> Yes.
Check {2,3,5} --> Yes.
Check {2,4,5} --> No.
This 4-itemset is out.
Therefore, {1,2,3,4} & {1,2,3,5}

9a. Draw an itemset lattice representing the data set given in **Table 5.22** . Label each node in the lattice with the following letter(s):

N: If the itemset is not considered to be a candidate itemset by the *Apriori* algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

F: If the candidate itemset is found to be frequent by the *Apriori* algorithm.

I: If the candidate itemset is found to be infrequent after support counting.



2. What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

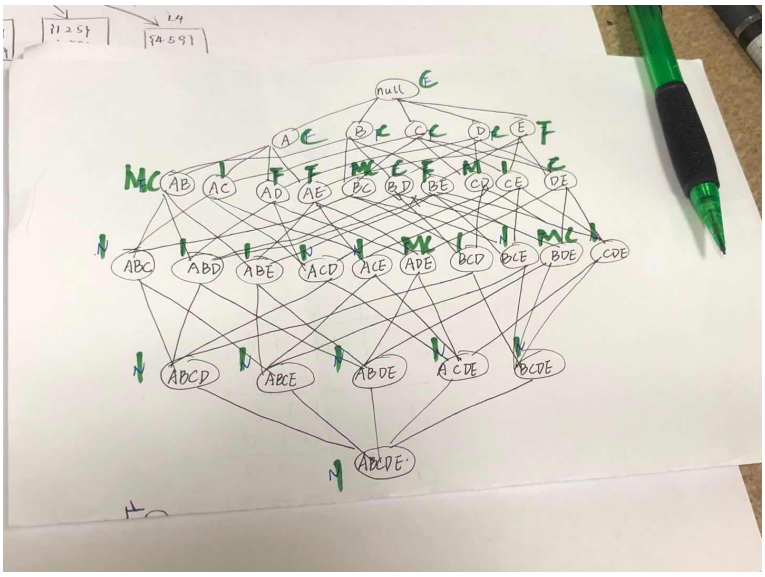
   Percentage of frequent itemsets = 16/32 = 50.0%(include the null set)

3. What is the pruning ratio of the *Apriori* algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

   Pruning ratio is the ratio of N to the total number of itemsets. Since the count of N = 11, therefore pruning ratio is 11/32 = 34.4%.

4. What is the false alarm rate (i.e., percentage of candidate itemsets that are found to be infrequent after performing support counting)?

   False alarm rate is the ratio of I to the total number of itemsets. Since the count of I = 5, therefore the false alarm rate is 5/32 = 15.6%.



13

a. Draw a contingency table for each of the following rules using the transactions shown in Table 5.23

Rules: {b}⟶{c}, {a}⟶{d}, {b}⟶{d}, {e}⟶{c}, {c}⟶{a}.

	c	c’
b	3	4
b’	2	1

	d	d’
a	4	1
a’	5	0

	d	d’
b	6	1
b’	3	0

	c	c’
e	2	4
e’	3	1

	a	a’
c	2	3
c’	3	2

b. Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

i. Support.

Rules	support	rank
b→c	.3	3
a→d	.4	2
b→d	.6	1
e→c	.2	4
c→a	.2	4

ii. Confidence.

Rules	Confidence	rank
b→c	3/7	3
a→d	4/5	2
b→d	6/7	1
e→c	2/6	5
c→a	2/5	4

iii. Interest (X→Y)=P(X, Y)P(X)P(Y).

Rules	interest	rank
b→c	.214	3
a→d	.72	2
b→d	.771	1
e→c	.167	5
c→a	.2	4

iv. IS(X→Y)=P(X,Y)P(X)P(Y).

Rules	IS	rank
b→c	.507	3
a→d	.596	2
b→d	.756	1
e→c	.365	5
c→a	.4	4

v. Klosgen(X→Y)=P(X, Y)×max(P(Y|X))−P(Y), P(X|Y)−P(X)), where P(Y|X)=P(X, Y)P(X).

Rules	Klosgen	rank
b→c	-0.039	2
a→d	-0.063	4
b→d	-0.033	1
e→c	-0.075	5
c→a	-0.045	3

vi. Odds ratio(X→Y)=P(X, Y)P(X¯, Y¯)P(X, Y¯)P(X¯, Y).

Rules	Odds Ratio	rank
b→c	.375	2
a→d	0	4
b→d	0	4
e→c	.167	3
c→a	.444	1

20

(a) For table I, compute support, the interest measure, and the φ correla- tion coefficient for the association pattern {A, B}. Also, compute the confidence of rules A → B and B → A.

s(A) = 0.1, s(B) = 0.9, s(A, B) = 0.09. I(A,B) = 9, φ(A,B) = 0.89.
c(A → B) = 0.9, c(B → A) = 0.9.

(b) For table II, compute support, the interest measure, and the φ correla- tion coefficient for the association pattern {A, B}. Also, compute the confidence of rules A → B and B → A.

s(A) = 0.9, s(B) = 0.9, s(A, B) = 0.89. I(A, B) = 1.09, φ(A, B) = 0.89.
c(A → B) = 0.98, c(B → A) = 0.98.

(c) What conclusions can you draw from the results of (a) and (b)?

Interest, support, and confidence are non-invariant while the φ-coefficient is invariant under the inversion operation. This is because φ-coefficient takes into account the absence as well as the presence of an item in a transaction.

2.1

The antecedents and consequents are shown above.
According to the code, we can easily find the max lift are rule 6 and 7. And the rules with the max confidence are Rule 46 & 48.

Rule 6: Antecedents = (CHILDRENS CUTLERY SPACEBOY), Consequents = (CHILDRENS CUTLERY DOLLY GIRL)
Rule 7: Antecedents = (CHILDRENS CUTLERY DOLLY GIRL), Consequents = (CHILDRENS CUTLERY SPACEBOY)

Rule 46: Antecedents = (SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS), Consequents = (SET/6 RED SPOTTY PAPER PLATES)
Rule 48: Antecedents = (SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED SPOTTY PAPER PLATES), Consequents = (SET/6 RED SPOTTY PAPER CUPS)

Because of difference between confidence and lift, the rule with the highest confidence is not as same as the rule with the highest lift.

Because condience is based on support of the itemsets, and lift is based on confidence of the rule and the support of the consequent.

2.2
These two items are asymmetric
{'Chocolate Coffee'} -> {'Chocolate Cake'} and {'Chocolate Cake'} -> {'Chocolate Coffee'} have the same Phi-Coefficient according to selection group