CS 412 Data Mining (22Sp): Assignment 3

Xu Ke (UIN: 675776713 NetID: kex5)

March. 13, 2022

NOTE: Unless otherwise specified, in this assignment, we use $sup(A)$ to denote the absolute support of an itemset A and $s(A)$ to denote the relative support of an itemset A.

1 Problem 1. True or False

Please justify your answers with **at most 3** sentences (1 point for true or false and 2 points for the justification).

1.1 Given two itemsets A and B, the range of χ^2 measure of A and B is [0,1], i.e., χ^2(A, B) $\in$ [0, 1].

$\boxed{False}$ According to the definition, we know that $\chi^2 = \frac{(Observed-Expected)^2}{Expected}$, thus there is no necessity that χ^2(A, B) should be in the range [0, 1] and it can far beyond 1 in case that the observed value is totally different from the expected value. (Lecture 4 Page 52)

1.2 Given two itemsets A and B, the lift $Lift(A, B)$ of A and B is null-invariant.

$\boxed{False}$ According to the definition, $Lift(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)}$ will result in range $[0, \infty]$, which implies the number of null transactions matters since it does change the measure value, thus the lift $Lift(A, B)$ of A and B should be not null-invariant. (Lecture 4 Page 54)

1.3 Given two frequent itemsets A and B, the lift $Lift(A, B)$ of A and B is significantly greater than 1 (i.e., $Lift(A, B) \gg 1$) implies that A and B are unlikely to happen together.

$\boxed{False}$ According to the property, $Lift(A, B) \gg 1$ implies that the itemsets A and B are greatly positively correlated, which means A and B are very likely to happen together. (Lecture 4 Page 51)

1.4 Given two itemsets A and B, the Allconf measure $Allconf(A, B)$ of A and B is null-invariant.

$\boxed{True}$ According to the definition, $Allconf(A, B) = \frac{s(A \cup B)}{max\{s(A), s(B)\}}$ will result in range $[0, 1]$, which implies the number of null transactions doesn't matter since it doesn't change the measure value, thus the Allconf $Allconf(A, B)$ of A and B should be null-invariant. (Lecture 4 Page 54)

1.5 Given two frequent itemsets A and B, we denote $s(A)$, $s(B)$ and $s(A \cup B)$ as the relative support of A, the relative support of B and the relative support of A$\cup$B, respectively. If $s(A \cup B) \ll s(A) \times s(B)$, it implies itemsets A and B frequently occur together.

$\boxed{False}$ According to the property, $Lift(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)}$, in this given case, we have $s(A \cup B) \ll s(A) \times s(B)$, which implies that $Lift(A, B) \ll 1$, thus itemsets A and B are greatly negatively correlated, which means A and B are very unlikely to occur together. (Lecture 4 Page 51)

1.6 Let A be an itemset and V be the set of all items. Is $A \subseteq V$ anti-monotone? If true, please explain the reason. Otherwise, please provide an counterexample.

$\boxed{True}$ According to the definition, a constraint c is anti-monotone if an itemset S violates constraint c, so does any of its supersets. In this given case, if $A \notin V$, which is the constraint, then any supersets of A would be definitely not in set V. (Lecture 5 Page 22)

1.7 Let A be an itemset, V be the set of all items, and $x \in V$ be an item. Is $x \in A$ succinct? If true, please explain the reason. Otherwise, please provide an counterexample.

$\boxed{True}$ According to the definition, succinctness means that if the constraint c can be enforced by directly manipulating the data. In this given case, we can just start with all items in V and remove all the transactions not in A. (Lecture 5 Page 32)

1.8 Let $V = \{a, b, c, d, e\}$ be the set of all items, $A = \{a, b\}$ and $B = \{a, b, c\}$ be two itemsets. Is $A \subseteq B$ monotone? If true, please explain the reason. Otherwise, please provide an counterexample.

$\boxed{False}$ According to the definition, a constraint c is monotone if an itemset S satisfies the constraint c, so does any of its supersets. In this given case, even though indeed $A \subseteq B$, however not any supersets would satisfy this constraint and an obvious counterexample would be $X = \{a, b, c, d\}$. (Lecture 5 Page 24)

2 Problem 2. Apriori Algorithm

The Apriori algorithm uses prior knowledge and follows Apriori pruning principle to mine frequent patterns.

2.1 Prove the correctness of Apriori pruning principle: *if there is any itemset which is infrequent, its superset should not even be generated. [Hint: you may prove it by contradiction]*

Let's prove the correctness of Apriori pruning principle by contradiction as follows:

First, let's assume that the Apriori pruning principle is not correct, which means if there is any itemset, let's say A, which is infrequent, its superset, let's say B, which satisfis $A \in B$ should be generated and to be frequent.

According to the definition, an itemset is frequent if the support of it is no less than a minsup threshold. Thus, in our assumption, the support of B must be equal or greater than the minsup threshold σ, $sup\{B\} \geq \sigma$.

Then, since the itemset B is the superset of A, that is to say the itemset A is the subset of B, which must satisfy that the support of A should be equal or greater than the support of B, $sup\{A\} \geq sup\{B\}$.

Thus, the support of A is also equal or greater than the threshold, $sup\{A\} \geq \sigma$. Thus with the definition of frequent itemset, itemset A should be frequent, which contradicts our assumption that itemset A is infrequent.

Hence, our assumption is not true, which implies the correctness of Apriori pruning principle.

2.2 Suppose we have a transaction database D that contains $|D|$ transactions. Let x be a frequent itemset of D, and x' be a nonempty subset of x whose relative support is s(x'). If we define the minimum support be minsup, what are the tight upper and lower bounds of the relative support of x (i.e. s(x))?

Since x' is the nonempty subset of x, thus all items belongs to x' must also belong to x, thus $\boxed{s(x) \leq s(x')}$.

Besides, since x is frequent, thus should satisfy threshold, $\boxed{s(x) \geq \frac{minsup}{|D|}}$.

2.3 Explain why Apriori algorithm is not efficient.

Apriori algorithm needs to scan the database multiple times to calculate the frequency of the itemsets in k-itemset. So, Apriori algorithm turns out to be very slow and inefficient, especially when memory capacity is limited and the number of transactions is large.

3 Problem 3. Null Invariance

Giving two itemsets A and B and the following contigency table (Table 1).

	A	$\neg A$	$\sum_{row}$
B	a	b	$a+b$
$\neg B$	c	d	$c+d$
$\sum_{col}$	$a+c$	$b+d$	$a+b+c+d$

Table 1: Contigency Table

3.1 What is the lift $Lift(A, B)$ of A and B? Under what condition would A and B be independent?

According to the definition of $Lift(A, B)$ (Lecture 4 Page 51), in this case we will have:

$$Lift(A,B) = \frac{s(A \cup B)}{s(A) \times s(B)} = \frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d} \times \frac{a+b}{a+b+c+d}} = \boxed{\frac{a \cdot (a+b+c+d)}{(a+c)(a+b)}} \tag{1}$$

From Lecture notes (Lecture 4 Page 51), we know that if A and B are independent, we should have $Lift(A, B) = 1$ and have condition $\boxed{\text{ad} = \text{bc}}$.

3.2 What is the imbalanced ratio $IR(A, B)$ of A and B?

According to the definition of $IR(A, B)$ (Lecture 4 Page 57) , in this case we will have:

$$IR(A,B) = \frac{|s(A) - s(B)|}{s(A) + s(B) - s(A \cup B)} = \frac{|\frac{(a+c)}{a+b+c+d} - \frac{(a+b)}{a+b+c+d}|}{\frac{(a+c)}{a+b+c+d} + \frac{(a+b)}{a+b+c+d} - \frac{a}{a+b+c+d}} = \boxed{\frac{|c-b|}{a+b+c}} \tag{2}$$

3.3 What is Kulczynski measure $Kulc(A, B)$ of A and B? And explain why $Kulc(A, B)$ is null-invariant.

According to the definition of $Kulc(A, B)$ (Lecture 4 Page 54), in this case we will have:

$$Kulc(A,B) = \frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)}) = \frac{1}{2}(\frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d}} + \frac{\frac{a}{a+b+c+d}}{\frac{a+b}{a+b+c+d}}) = \boxed{\frac{1}{2}(\frac{a}{a+c} + \frac{a}{a+b})} \tag{3}$$

In this given case,since the result of $Kulc(A, B)$ only relates to a,b and c but has no connection with the null transaction d, which implies that $Kulc(A, B)$ is null-invariant.

3.4 What is the difference between lift $Lift(A, B)$ and cosine measure $Cosine(A, B)$? Why would this difference make $Cosine(A, B)$ null-invariant?

According to the definition of $Lift(A, B)$ and $Cosine(A, B)$, we know that:

$$Lift(A,B) = \frac{s(A \cup B)}{s(A) \times s(B)} = \frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d} \times \frac{a+b}{a+b+c+d}} = \frac{a \cdot (a+b+c+d)}{(a+c)(a+b)} \tag{4}$$

$$Cosine(A,B) = \frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}} = \frac{\frac{a}{a+b+c+d}}{\sqrt{\frac{a+c}{a+b+c+d} \times \frac{a+b}{a+b+c+d}}} = \frac{a}{\sqrt{(a+c)(a+b)}} \tag{5}$$

From the two equations above, it is clear that $Lift(A, B)$ has (a+b+c+d) in the numerator while $(a+b+c+d)^2$ in the denominator due to $s(A) \times s(B)$. $Cosine(A, B)$ has (a+b+c+d) in both the numerator and the denominator due to $\sqrt{s(A) \times s(B)}$, thus this term related to d can be cancelled. Thus, $Lift(A, B)$ depends on null transaction d and thus null-variant, but $Cosine(A, B)$ has no connection with null transaction d and thus null-invariant.

4 Problem 4. Constraint-Based Pattern Mining

Suppose you have a transaction database from a grocery store shown in Table 2 and the corresponding product table in Table 3.

TID	Transaction
10	a, b, c, d, f, h
20	a, c, d, e, g, h
30	a, c, e, g, h
40	b, d, e, f, g
50	a, b, e, d, f, g

Table 2: Transaction Database

Item	Price	Profit
a	10	4
b	16	8
c	46	20
d	40	0
e	37	12
f	30	-10
g	45	-5
h	100	-20

Table 3: Product Table

4.1 Given a collection of constraints as follows, identify their types (monotone, anti-monotone, data anti-monotone, succinct, convertible). If multiple types coexist in the following constraints, please list them all.

Monotone: If an itemset S satisfies the constraint c, so does any of its supersets.
Anti-monotone: If an itemset S violates constraint c, so does any of its supersets.
Data anti-monotone: In the mining process, if a data entry t cannot contribute to a pattern p satisfying c, t cannot contribute to p's superset either.
Succinct: If the constraint c can be enforced by directly manipulating the data.
Convertible: c can be converted to monotonic or anti-monotonic if items can be properly ordered in processing

4.1.1 Transaction that includes the item a.

It is $\boxed{monotone}$, $\boxed{succinct}$ and $\boxed{dataanti - monotone}$. With the definitions shown above, since any superset of one set which contains item a must also include a, thus monotone. Besides, we can enforce this constraint by deleting all transactions that don't contain item a, thus succinct. If a transaction doesn't contribute to a pattern p containing item a, then it cannot contribute to p's any superset, thus data anti-monotone.

4.1.2 Total price of all purchased items is less than $250.

It is $\boxed{anti - monotone}$. With the definitions shown above, since price is always positive and thus if the current total price of all purchased items is lager or equal than $250, then clearly all its supersets will also lager or equal than $250, thus anti-monotone.

4.1.3 Total price of the transaction is at least $200.

It is $\boxed{monotone}$ and $\boxed{dataanti-monotone}$. With the definitions shown above, since price is always positive and thus if the current total price of the transaction is at least $200 then clearly the total price of the transactions in any supersets will also be at least $200, thus monotone. If a transaction doesn't contribute to a pattern p whose total price of the transaction is at least $200, then it cannot contribute to p's any superset, thus data anti-monotone.

4.1.4 Total price of all purchased items is less than $50.

It is $\boxed{anti-monotone}$. With the definitions shown above, since price is always positive and thus if the current total price of all purchased items is lager or equal than $50, then clearly all its supersets will also lager or equal than $50, thus anti-monotone.

4.1.5 The minimal price of the transaction is less than $50.

It is $\boxed{monotone}$, $\boxed{succinct}$ and $\boxed{dataanti-monotone}$. With the definitions shown above, since if current the minimal price of the transaction is less than $50 then the minimal price of the any supersets will also be less than $50. Besides, we can start with only items whose price is less than $50 and remove transactions with high-price items only, thus succinct. If a transaction doesn't contribute to a pattern p whose smallest item is less than $50, then it cannot contribute to p's any superset, thus data anti-monotone.

4.2 Identify the type of the constraint sum(S.price) >100, and justify your answer.

It is $\boxed{monotone}$ and $\boxed{dataanti-monotone}$. With the definitions shown above, since price is always positive and thus if the current total price of the transaction is larger than 100 then clearly the total price of the transactions in any supersets will also be gretaer than 100, thus monotone. If a transaction doesn't contribute to a pattern p whose total price of the transaction is larger than $100, then it cannot contribute to p's any superset, thus data anti-monotone.

5 Problem 5. Multi-level Pattern Mining

Given the association rules below, answer the following questions.

Rule (5.1) *buys(X, "laptop")* $\Longrightarrow$ *buys(X, "HP printer")*
[support = 8%, confidence = 70%]

Rule (5.2) *age(X, "18...25")* $\wedge$ *occupation(X, "student")* $\Longrightarrow$ *buys(X, "laptop")*
[support = 5%, confidence = 84%]

Rule (5.3) *buys(X, "HP laptop")* $\Longrightarrow$ *buys(X, "HP printer")*
[support = 7%, confidence = 72%]

Rule (5.4) *age(X, "18...25")* $\wedge$ *buys(X, "HP laptop")* $\Longrightarrow$ *buys(X, "HP printer")*
[support = 2%, confidence = 60%]

5.1 Identify which association rule (single-dimensional rule, multi-dimensional rule) is used. If it is multi-dimensional rule, please specify whether it is inter-dimension association rule or hybrid-dimension association rule.

5.1.1 What association rule is Rule (5.1)?

It uses $\boxed{single-dimensionalrule}$ since items are all in "buys" dimension.

5.1.2 What association rule is Rule (5.2)?

It uses $\boxed{multi-dimensionalrule}$ since items in 2 dimensions or predicates. It is $\boxed{inter-dimensionassociationrule}$ since there is no repeated predicates.

5.1.3 What association rule is Rule (5.4)?

It uses $\boxed{multi - dimensionalrule}$ since items in 2 dimensions or predicates. It is $\boxed{hybrid - dimensionassociationrule}$ since there is repeated predicates.

5.2 If Rules (5.1) and (5.3) are both mined, is Rule (5.3) or Rule (5.1) redundant? Why?

$\boxed{(Rule5.3)}$ is redundant. According to the definition, a rule is redundant if its support is close to the "expected" value, according to its "ancestor" rule, and it has a similar confidence as its "ancestor". In this given case, Rule (5.3) should be able to be "derived" from Rule (5.1), which implies that Rule (5.1) is the "ancestor" of the Rule (5.3) and we can tell that their support and confidence are very similar.