Quiz 3

- There are 5 problems total worth 12 points as shown in each question.
- You must not communicate with other students during this test.
- No books, notes allowed.
- No other electronic device except calculators are allowed. You cannot use your mobile as calculators.
- This is a 30 minute exam.
- Do not turn this page until instructed to.
- There are several different versions of this exam.

1. Fill in your information:

Full Name: ______________________________

NetID: ______________________________

1/1. (1 point) Why is the runtime comparable for large support in the Apriori and FP-growth algorithms?

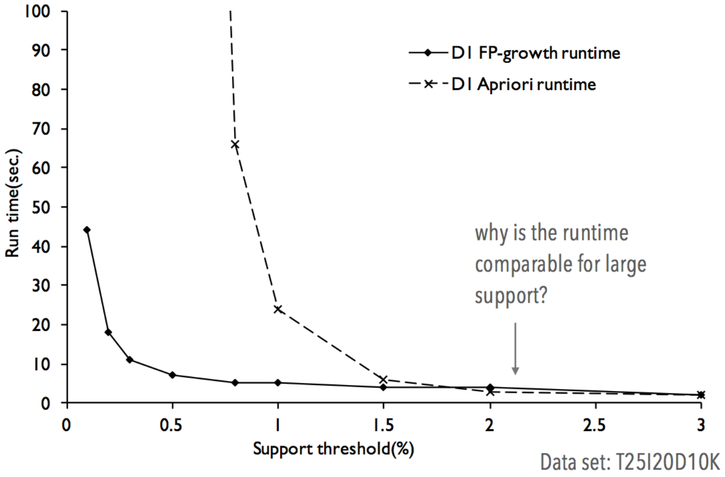

Figure 1: FP-growth v.s. Apriori: scalability with the support threshold.

Solution. When we raise the support, even at a small percentage, the number of candidates generated shrinks exponentially. FP-growth works better than Apriori under the condition that the number of candidates generated grows exponentially. Higher support removes this condition. FP-growth and Apriori perform same for large support.

2/1. (3 points) Below is the contingency table that summarizes a supermarket transaction data of hot dogs and hamburgers sell.

	$hotdogs$	$\overline{hotdogs}$	$\sum_{col}$
$hamburgers$	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
$\sum_{row}$	3000	2000	5000

(a) Suppose that the association rule ” hot dogs ⇒ hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50% is this association rule strong?

(b) Based on the given data, is the purchase of hot dogs independent of that of humburgers? If not, what kind of correlation relationship exists between the two?

(c) Calculate and compare two measures of correlation: *Lift* and *Kulczynski* on the above given data.
(Hint: $Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}, \quad Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$.)

Solution.

A. Support = 40% > 25%. Confidence = 66.7% > 50% . Rule is strong.

B. Positive correlation

C. Lift = 1.33 , Kulc = 0.733

2/2. (3 points) Below is the contingency table that summarizes a supermarket transaction data of hot dogs and hamburgers sell.

	$hotdogs$	$\overline{hotdogs}$	$\sum_{col}$
$hamburgers$	80	100	180
$\overline{hamburgers}$	120	700	820
$\sum_{row}$	200	800	1000

(a) Suppose that the association rule ” hot dogs ⇒ hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50% is this association rule strong?

(b) Based on the given data, is the purchase of hot dogs independent of that of humburgers? If not, what kind of correlation relationship exists between the two?

(c) Calculate and compare two measures of correlation: *Lift* and *Kulczynski* on the above given data.
(Hint: $Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}, \quad Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$.)

Solution.

A. Support = 8% < 25%. Confidence = 40% < 50% . Rule is not strong.

B. Negative correlation

C. Lift = 2.2222 , Kulc = 0.4222

3/1. (3 points) Suppose a WalMart manager is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. For the following cases, state the characteristics (i.e., categories) of every constraint in each case.

A. The average price of all the items in each pattern is greater than \$50.

B. The sum of the price of all the items with profit over \$5 in each pattern is at least \$200.

Solution.

A. $C : avg(S.price) > 50$. This constraint is strongly convertible.

(a) It can be converted to anti-monotonic if the items are sorted in descending order of their prices.

(b) It can be converted to monotonic if the items are sorted in ascending order of their prices.

B. $C1 : min(S.profit) > 5$ is succinct, or data anti-monotone. $C2 : sum(S.price) \geq 200$ is monotone.

3/2. (3 points) Suppose a WalMart manager is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. For the following cases, state the characteristics (i.e., categories) of every constraint in each case and how to mine such patterns most efficiently

A. The price difference between the most expensive item and the cheapest one in each pattern must be within \$100.

B. The sum of the profit of all the items in each pattern is above \$10 and each such item is priced over \$10.

Solution.

A. $C : range(S.price) \leq \$100$ is antimontonic.

B. $C1 : min(S.price) > \$10$ is succinct, or data anti-monotone. $C2 : sum(S.profit) \geq \$10$ is convertible monotone

---

4/1. (3 points) Below are some transaction patterns with their supports from a transaction dataset. Identify all the core patterns ($\boldsymbol{\tau} = \mathbf{0.3}$) for each given transaction pattern. Please show your calculation steps briefly.

Transaction pattern	Support (number of transactions)
ab	200
abc	180
abe	200
bcde	160

Solution.

Transaction pattern	Support (number of transactions)
ab	a,ab
abc	a, c, ab, ac, bc,abc
abe	a, e, ab, ae, be, abe
bcde	c, d, e, bc, bd, be, cd, ce, de, bcd, bce, bde, cde, bcde

---

4/2. (3 points) Below are some transaction patterns with their supports from a transaction dataset. Identify all the core patterns ($\boldsymbol{\tau} = \mathbf{0.5}$) for each given transaction pattern. Please show your calculation steps briefly.

Transaction pattern	Support (number of transactions)
ad	200
abc	160
bd	200
bcde	180

Solution.

Transaction pattern	Support (number of transactions)
ad	a,ad
abc	ab, ac,abc
bd	bd
bcde	c, e, bc, be, cd, ce, de, bcd, bce, bde, cde, bcde

5/1. (2 points) Below is a transaction dataset and item prices of each transaction. Assume each kind of item has one distinct positive price. Use the FP-Growth algorithm to find all the frequent patterns that satisfy the requirements **min_sup = 2** and **min{S.price}** $\leq$ **2**.

TID	Items (Price)
I1	1, 2, 4
I2	2, 3, 5
I3	1, 2, 3, 5
I4	2, 5

Please write out each projected database that you create and mark those items that are pruned at each step.

Solution.

A. 4 is pruned

B. Ordered Item List : $< 2 : 4 >, < 5 : 3 >, < 3 : 2 >, < 1 : 2 >$

C. Frequent patterns : $< 1 : 2 >, < 12 : 2 >, < 23 : 2 >, < 235 : 2 >, < 25 : 3 >, < 2 : 4 >$

5/2. (2 points) Below is a transaction dataset and item prices of each transaction. Assume each kind of item has one distinct positive price. Use the FP-Growth algorithm to find all the frequent patterns that satisfy the requirements **min_sup = 2** and **min{S.price}** $\leq$ **2**.

TID	Items (Price)
I1	1, 2, 3, 5
I2	2, 4
I3	1, 2, 4, 5
I4	2, 5

Please write out each projected database that you create and mark those items that are pruned at each step.

Solution.

A. 3 is pruned

B. Ordered Item List : $< 2 : 4 >, < 5 : 3 >, < 1 : 2 >, < 4 : 2 >$

C. Frequent patterns : $< 24 : 2 >, < 1 : 2 >, < 12 : 2 >, < 15 : 2 >, < 125 : 2 >, < 25 : 3 >, < 2 : 4 >$
