CSE 40647/60647: Data Science **Fall 2019**

Homework 4: Written Assignments

Handed Out: November 14, 2019 *Due: December 6, 2019 11:55pm*

Save your homework submission as *NETID-hw4-written.pdf*.

1 FP-Growth (30 points)

A database has 10 transactions. Let $min\_sup = 2$. Items are a, b, c, d, and e.

Trans. ID	Itemset
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

Draw the first FP-tree that the FP-Growth algorithm creates when given this transaction database. By saying the "first", this FP-tree should not be a conditional FP-tree. Use FP-Growth to find all the frequent patterns and their support. Attach the FP-tree (either typed or hand-written+scanned) and write down the patterns and support in your PDF.

2 Pattern Evaluation Measures (10 points)

The definitions of two measures, *lift* and *cosine*, look rather similar as shown below,

$$\text{lift}(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)}, \tag{1}$$

and

$$\text{cosine}(A, B) = \frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}, \tag{2}$$

where $s(X)$ is the *relative* support of itemset X. Which measure is *null-invariant*, and which is not, and why? Can you prove it? You must formally define what is null-invariant using the symbols and give your proof.

3 Closed Patterns (20 points)

A database has 4 transactions as shown below. Let $min\_sup = 2$. Items are A, B, C, D, E, F, and G.

Trans. ID	Itemset
1	{A, C, F, G}
2	{A, B, C, F}
3	{A, B, C, D, F}
4	{B, D, E}

Which patterns from the following are **closed patterns**? Please briefly describe your idea for each pattern on why it is closed or not.

- Pattern 1: {D}
- Pattern 2: {A, B, C, F}
- Pattern 3: {B, F}
- Pattern 4: {B, D}
- Pattern 5: {A, C, F}

4 Sequential Patterns (20 points)

A sequence database has 3 sequences as shown below. Items in the same parenthesis means they were got together in one event. Let $min\_sup = 2$. Items are A, B, C, D, F, and G. Which patterns from the following are **sequential patterns**? Please briefly describe your idea for each pattern on why it is a good sequential pattern or not.

Seq. ID	Sequence
1	(AB)C(FG)G
2	(AD)CB(ABF)
3	AB(FG)

- Pattern 1: ACF
- Pattern 2: (FG)B
- Pattern 3: (FG)
- Pattern 4: B(FG)
- Pattern 5: GF

CSE 40647/60647: Data Science **Fall 2019**

Homework 4: Programming Assignments

Handed Out: November 12, 2019 *Due: December 6, 2019 11:55pm*

Save your homework submission as *NETID-hw4-programming.zip*. The zip file has one pdf file *NETID-hw4-programming.pdf*, one code file, the Dataset-apriori.txt file, and one README file.
In the README file, please specify the python version you used and how to run your code in command line.

Apriori (50 points)

Please use **Python** to solve the problem. You are NOT allowed to directly call any frequent pattern mining functions (like the Apriori functions in Scikit).

A database has 10 transactions. Let $min\_sup = 2$. Items are a, b, c, d, and e.

Trans. ID	Itemset
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

Use Python to implement Apriori to find all frequent patterns (i.e., frequent itemsets) and their counts from the transaction database.

Output: Write down the patterns and their support in the pdf. Save your code as NETID-hw4.py.