CSE 40647/60647: Data Science **Fall 2017**

Homework 3

Handed Out: September 21, 2017 *Due: October 03, 2017 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. You can handwrite trees/figures and scan them into PDF. Name your pdf file as YourNetid-HW3.pdf
- You need to explain the logic of your answer/result for every question. A result/answer without any explanation will not receive any point.
- It is OK to discuss the problems with the TA and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the Honor code on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

2 Question 1 (40 points)

A database has 10 transactions. Let $min\_sup = 2$.

trans_id	items
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 }

1. (20′) Please use Apriori Algorithm to find all frequent patterns and their counts. (Note: Please list frequent patterns and their counts step by step. Only results are not acceptable)

2. (20′) Please use FP-growth Algorithm to find all frequent patterns and their counts. (Note: You will need to draw the FP-Tree and list frequent patterns and their counts step by step. Only results are not acceptable.)

Solution:

1. (20′) **Answer:**
   1-itemset candidates C1: a: 8; b: 7; c: 6; d: 5; e: 3.
   Compare C1 with minimum support. *The frequent 1-itemsets F1:*
   a: 8; b: 7; c: 6; d: 5; e: 3.

   2-itemset candidates C2: ab: 5; ac: 4; ad: 4; ae: 2; bc: 5; bd: 3; be: 0; cd: 3; ce: 2; de: 2.
   Compare C2 with minimum support. *The frequent 2-itemsets F2:*
   ab: 5; ac: 4; ad: 4; ae: 2; bc: 5; bd: 3; cd: 3; ce: 2; de: 2.

   3-itemset candidates C3: abc: 3; abd: 2; bcd: 2; cde: 1; acd: 2; ace: 1; ade: 2.
   Compare C3 with minimum support. *The frequent 3-itemsets F3:*
   abc: 3; abd: 2; bcd: 2; acd: 2; ade: 2.

   There is no frequent 4-itemsets.

2. (20′) **Answer:**
   The FP-Tree is shown below in Fig. 1 (Dashed links between nodes of same labels are not required)

Frequent itemsets found (descending order by frequency of each item):

Item	Frequent Patterns
a	**a**
b	**b**, **ba** (from *b*-conditional FP-tree)
c	**c**, **cb** and **ca** (from *c*-conditional FP-tree), **cba** (from *cb*-conditional FP-tree)
d	**d**, **dc** and **db** and **da** (from *d*-conditional FP-tree), **dcb** and **dca** (from *dc*-conditional FP-tree), **dba** (from *db*-conditional FP-tree)
e	**e**, **ed** and **ec** and **ea** (from *e*-conditional FP-tree), **eda** (from *ed*-conditional FP-tree)

Conditional FP-trees can easily be derived from Fig. 1 step by step.

3 Question 2 (20 points)

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

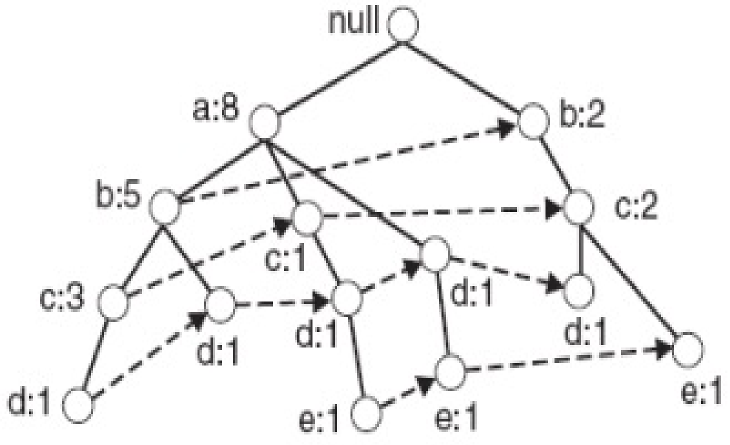

Figure 1: FP tree.

$\textit{lift}(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)}$

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

where $s(A)$ is the *relative* support of itemset A. Explain why one of these two measures is *null-invariant* but the other is not.

Solution:

1. (20′) **Answer:**
   A measure is null-invariant if the value of the measure does not change with the number of null-transactions.
   cosine is null-invariant while *lift* is not.
   Let n be the total number of transactions, and $count(\neg(A \cup B))$ be the number of null-transactions.
   $\textit{lift}(A, B) = \frac{s(A \cup B)}{s(A) \times s(B)} = \frac{\frac{count(A \cup B)}{n}}{\frac{count(A)}{n} \times \frac{count(B)}{n}} = \frac{count(A \cup B) \times n}{count(A) \times count(B)} = \frac{count(A \cup B) \times (count(A \cup B) + count(\neg(A \cup B))}{count(A) \times count(B)}$
   $\textit{cosine}(A, B) = \frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}} = \frac{\frac{count(A \cup B)}{n}}{\sqrt{\frac{count(A)}{n} \times \frac{count(B)}{n}}} = \frac{count(A \cup B)}{count(A) \times count(B)}$
   We can clearly see that *cosine* is invariant with the number of null-transactions, while *lift* is not.

4 Question 3 (20 points)

A database has 4 transactions. Let $min\_sup = 2$.

trans_id	items
1	{A, C, F, G}
2	{A, B, C, F}
3	{A, B, C, D, F}
4	{B, D, E}

Please choose the closed patterns from the following patterns. (Note: Briefly describe your idea on how to decide which is closed pattern and which is not.)

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

Solution:

1. (20') **Answer:** X is a closed frequent itemset in a data set S if there exists no proper super-item(super-pattern) Y such that Y has the same support count as X in S, and X satisfies minimum support.

   The closed patterns in the lists are: Pattern 2 (support = 2), Pattern 4 (support = 2), Pattern 5 (support = 3).

   {B, D} is a super-pattern of Pattern 1 {D}. Their supports are both 2. Pattern 1 is not closed.

   Pattern 2 is a super-pattern of Pattern 3, and their supports are both 2. Pattern 3 is not closed.

5 Question 4 (20 points)

A sequence database has 3 sequences. Items in the same parenthesis means they were got together in one event. Let $min\_sup = 2$.

sequence_id	sequence
1	(AB)C(FG)G
2	(AD)CB(ABF)
3	AB(FG)

Please choose the sequential patterns from the following patterns. (Note: Briefly describe your idea on how to decide which is sequential pattern and which is not.)

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

Solution:

1. (20′) **Answer:** Traditionally, sequential pattern mining is being used to find subsequences that appear often in a sequence database, those subsequences are called the frequent sequential patterns.

   The sequential patterns in the lists are: Pattern 1 (sup = 2, Seq. 1 and 2), Pattern 3 (sup = 2, Seq. 1 and 3) and Pattern 4 (sup = 2, Seq. 1 and 3).

   Pattern 2 (sup = 0) and Pattern 5 (sup = 0) are not sequential patterns. Note that Pattern 5 has two single-item events: It is not (GF), i.e., Pattern 3.

Patterns	Support
A	3
AG	2
AF	2
AC	2
ACF	2
AB	2
ABF	2
(AB)	2
B	3
BG	2
BF	3
B(FG)	2
C	2
CF	2
F	3
(FG)	2
G	2

6 Follow-up Questions on Project (0 point)

Your answers may be open to the class by the instructor. We encourage communications between teams. Your answers may be released to graders.

Your partner name: ______________________________ (NetID: ___________________________)
Write "N/A" if you will do the project by your own.

1. Task 1: Data cleaning and integration

   a) How many unique papers and how many unique authors are there in your integrated and cleaned dataset?

   b) Find "matrix" experts: List the top three authors who published at least 3 papers AND used the word "matrix" the most frequently in their papers (i.e., the highest average number of "matrix" in their publications).

   c) Find "long-title" authors: List the top three authors who published at least 3 papers AND preferred long titles in their papers (i.e., the highest average length of paper titles).

2. Task 2: Entity name recognition

   d) How many unique case-insensitive entity names (like "support vector machines") have you discovered in the dataset? List the top 20 entity names and their support (i.e., the number of papers that have at least one such entity name) if you have.

   e) Briefly explain your technique(s).

7 Mid-semester Survey (0 point)

`https://www.surveymonkey.com/r/G32K2PT`