CS550: Massive Data Mining and Learning Spring 2018
Problem Set 1
Due 11:59pm Sunday, February 25, 2018

Only one late period is allowed for this homework (11:59pm Monday 2/26)

Submission Instructions

Assignment Submission: Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

Late Day Policy: Each student will have a total of **two** free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

Honor Code: Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Honor Code.

(Signed) ____FT____

If you are not printing this document out, please type your initials above.

Answer to Questions 1

ii)During the mapping stage:

I checked every user and pair their existed friends. The structure of the key-value pair is (ID, (possible suggested friend,1)). The existed friends are grouped in the structure (ID,(existed friend, -1)).

While during reducing, all the pairs with same keys will be put together and then I looped the all values and maintain a hash map to count the same ids of suggested friends, omit the existed friends. Then I sorted the map and output the top ten recommendations.

iii)

user	Re1	Re2	Re3	Re4	Re5	Re6	Re7	Re8	Re9	Re10
924	439	2409	6995	11860	15416	43478	45881			
8491	8493	8488	8489	8492	8494	8499	8501	8503	8504	
8492	8939	8940	8943	8944						
9019	9022	317	9023							
9020	9021	9016	9017	9022	317	9023				
9021	9020	9016	9017	9022	317	9023				
9022	9019	9020	9021	317	9016	9017	9023			
9990	13134	13478	13877	34299	34485	34642	37941			
9992	9987	9989	35667	9991						
9993	9991	13134	13478	13877	34299	32285	34642	37941		

Answer to Questions 2(a)

When A and B are independent, $conf(A \to B) = Pr(B|A) = \frac{Pr(AB)}{Pr(A)} = \frac{Pr(A) \cdot Pr(B)}{Pr(A)} = Pr(B)$, which means that $conf(A \to B)$ is only determined by $Pr(B)$. Therefore, A could be seen as frequent when $Pr(B)$ is high though they are independent. This may lead to a wrong conclusion.

Because $S(B) = \frac{support(B)}{N} = Pr(B)$, we know $lift(A \to B)$ and $conv(A \to B)$ take $Pr(B)$ into account.

Answer to Questions 2(b)

$$conf(A \to B) = Pr(B|A) = \frac{Pr(AB)}{Pr(A)} \neq conf(B \to A) = Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

$$lift(A \to B) = \frac{conf(A \to B)}{S(B)} = \frac{Pr(AB)}{Pr(A)Pr(B)} = \frac{Pr(AB)}{Pr(B)\,Pr(A)} = \frac{conf(B \to A)}{S(A)} = lift(B \to A)$$

$$conv(A \to B) = \frac{1 - S(B)}{1 - conf(A \to B)} = \frac{1 - Pr(B)}{1 - \frac{Pr\ (AB)}{Pr\ (A)}} = \frac{Pr\ (A) - Pr\ (A)Pr(B)}{Pr\ (A) - Pr\ (AB)} \neq conv(B \to A)$$

Therefore, only lift is symmetric.

For example, if we have four baskets*, A, B, AB* and *BC.*
Then $conf(A \to B) = P(B|A) = \frac{1}{2}$, $conf(B \to A) = Pr(A|B) = \frac{1}{3}$
$S(B) = \frac{3}{4}, S(A) = \frac{1}{2}$
$lift(A \to B) = \frac{conf(A\to B)}{S(B)} = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}$, $lift(B \to A) = \frac{conf(B\to A)}{S(A)} = \frac{1}{3} \cdot \frac{2}{1} = \frac{2}{3}$
$conv(A \to B) = \frac{1-S(B)}{1-conf(A\to B)} = \frac{1-\frac{3}{4}}{1-\frac{1}{2}} = \frac{1}{2}$, $conv(B \to A) = \frac{1-S(A)}{1-conf(B\to A)} = \frac{1-\frac{1}{2}}{1-\frac{1}{3}} = \frac{3}{2}$

Answer to Questions 2(c)
If A and B occur simultaneously every time, then
$conf(A \to B) = 1$.
$lift(A \to B) = \frac{1}{S(B)}$, it is determined by $S(B)$.
$conv(A \to B) = \frac{1-S(B)}{1-conf(A\to B)} \to +\infty$.
Therefore, confidence and conviction are desirable.

For example, we have three baskets, AB, CD, CDE where A occurs every time B occurs and C occurs every time D occurs.
$lift(A \to B) = \frac{1}{S(B)} = \frac{N}{support(B)} = \frac{3}{1} = 3$, $lift(C \to D) = \frac{1}{S(D)} = \frac{N}{support(D)} = \frac{3}{2}$
Though, these two rules are both 100% rules, they have different lift scores.

Answer to Questions 2(d)

DAI93865 -> FRO40251	1.0
GRO85051 -> FRO40251	0.999176276771005
GRO38636 -> FRO40251	0.9906542056074766
ELE12951 -> FRO40251	0.9905660377358491
DAI88079 -> FRO40251	0.9867256637168141

Answer to Questions 2(e)

(DAI23334,ELE92920)->DAI62779	1.0
(DAI31081,GRO85051)->FRO40251	1.0
(DAI55911,GRO85051)->FRO40251	1.0
(DAI62779,DAI88079)->FRO40251	1.0
(DAI75645,GRO85051)->FRO40251	1.0

Answer to Questions 3(a)

If a column has m 1's, the number of columns with m 1's of n is C_n^m.

The number of these columns without 1 in one of k chosen rows is equal to the number that all m columns are not chosen C_{n-k}^{m}

The probability of no 1 in the k chosen rows is p= $\frac{C_{n-k}^{m}}{C_{n}^{m}} = \frac{\frac{(n-k)!}{m!(n-k-m)!}}{\frac{n!}{m!(n-m)!}} = \frac{(n-m)!(n-k)!}{n!(n-k-m)!} = \frac{(n-k)!}{n!}\frac{(n-m)!}{(n-m-k)!} \leq \left(\frac{(n-k)}{n}\right)^{m}$

Answer to Questions 3(b)

$$\left(\frac{(n-k)}{n}\right)^{m} = \left(1-\frac{k}{n}\right)^{m} = \left[\left(1-\frac{k}{n}\right)^{\frac{n}{k}}\right]^{\frac{mk}{n}} = e^{-\frac{mk}{n}} \leq e^{-10}$$

Therefore, $\frac{mk}{n} \leq 10, \rightarrow k \leq \frac{10n}{m}$

Answer to Questions 3(c)

For example, the matrix with two columns is $\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$. The Jaccard Similarity is $J = 0.5$

When the second row and the third row are chosen as the random row r to be put at the first row, their minhash values are same. The probability is 0.5.