CS550: Massive Data Mining
Homework 1

Due 11:59pm Monday, February 21, 2022
Please see the homework file for late policy

Submission Instructions

Honor Code Students may have discussions about the homework with peers. 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 with whom they have discussions about the homework. Directly using the code or solutions obtained from the web or from others is considered an honor code violation. We check all the submissions for plagiarism and take the honor code seriously, and we hope students to do the same.

Discussions (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) SK*____

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

Answer to Question 1

(1) Code for "People You Might Know" is in a file named *question_1.py*

(2) Description of Algorithm:

Firstly, we are using map function to get pairs with different x, where x represents type of connection. x = 0 represents direct friendship while x = 1 indicates that pair share a mutual friend.

Now we are removing pairs who are already friends. We are basically getting rid of all the pairs where value of x is 0.

In addition, we are adding the number of mutual friends for each pair. We have to give recommendations in descending order of number of mutual friends. We are using map function to make tuples from both sides of users so that we can sort it in descending order properly. After we have all these pairs, we will just group the recommendations of every user and sort them in descending order.

(3) Recommendation for the user with the following userIDs:

- **924** 439,2409,6995,11860,15416,43748,45881
- **8941** 8943,8944,8940
- **8942** 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** 9991,13134,13478,13877,34299,34485,34642,37941

Answer to Question 2(a)

Confidence is defined as the probability of occurrence of B in the basket if the basket already contains A. It is ignoring the probability of occurrence of B which is a drawback because it can give you incorrect results. There are some items with high support like bread or milk that will produce large confidence values with many items. We will have a lot of association rules having these items which are not useful.

Lift and Conviction don't suffer from this drawback because they take the probability of B into consideration as we can see in their formula.

Answer to Question 2(b)

Lift is symmetrical and Confidence and Conviction are not symmetrical.

- **Confidence** It is given by,
  $conf(A \to B) = Pr(B|A) = \frac{Pr(AB)}{P(A)}$
  $conf(B \to A) = Pr(A|B) = \frac{Pr(AB)}{P(B)}$
  As P(A) is not equal to P(B),$conf(A \to B)$ is not equal to $conf(B \to A)$
  Therefore, Confidence is not symmetrical.

- **Lift** Lift is given by,
  $lift(A \to B) = \frac{conf(A \to B)}{S(B)} = \frac{S(AB)}{S(A)S(B)}$
  $lift(B \to A) = \frac{conf(B \to A)}{S(A)} = \frac{S(AB)}{S(A)S(B)}$
  As both terms are equal, Lift is symmetrical.

- **Conviction** Conviction is not symmetrical and we can prove that using counterexample.
  Let's say we have the following baskets: AB, A, BC, AD
  We have,
  S(A) = $\frac{3}{4}$
  S(B) = $\frac{1}{2}$
  $Pr(A \cap B) = \frac{1}{4}$
  $conv(A \to B) = \frac{1-S(B)}{1-conf(A \to B)} = \frac{1-S(B)}{1-\frac{Pr(AB)}{S(A)}} = \frac{1-1/2}{1-\frac{1/4}{3/4}} = \frac{1/2}{2/3} = 3$

  $conv(B \to A) = \frac{1-S(A)}{1-conf(B \to A)} = \frac{1-S(A)}{1-\frac{Pr(AB)}{S(B)}} = \frac{1-3/4}{1-\frac{1/4}{1/2}} = \frac{1/4}{1/2} = \frac{1}{2}$
  As we can see that conviction is not symmetrical.

Answer to Question 2(c)

As written earlier, $conf(A \rightarrow B) = Pr(AB)/P(A)$ and confidence gets its maximum value when Pr(AB) = P(A). The maximum value of confidence will be 1 when the above condition holds. Therefore, **Confidence is a desirable measure.**

Lift depends on the value of Pr(B) and the value of lift may or may not be maximal and can vary. Therefore, **Lift is not a desirable measure.**

When confidence gets its maximum value, i.e. 1, a conviction will be infinity which is the maximum value. Therefore, **Conviction is a desirable measure.**

Answer to Question 2(d)

Code is in file named *question_2.py*

Associate Rules	Confidence
DAI93865 → FRO40251	1.0
GRO85051 → FRO40251	0.999176276771005
GRO38636 → FRO40251	0.9906542056074766
ELE12951 → FRO40251	0.9905660377358491
DAI88079 → FRO40251	0.9867256637168141

Answer to Question 2(e)

Code is in file named *question_2.py*

Associate Rules	Confidence
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 Question 3(a)

Given, column has,
m $\rightarrow 1s$
$n \rightarrow 0s$
To prove:-
P(Getting "don't know" as min-hash value for this column) is at most $(\frac{n-k}{n})^m$

The number of different combinations with m 1's out of n is $\binom{n}{m}$
In any of the k selected rows, number of columns that do not have 1 is $\binom{n-k}{m}$.

P(Getting "don't know" as min-hash value for this column) is given by,
$= \binom{n-k}{m} / \binom{n}{m}$
$= \frac{(n-k)!}{m!(n-k-m)!} * \frac{m!(n-m)!}{n!}$
$= (\frac{n-k}{n})(\frac{n-k-1}{n-1})...(\frac{n-k-m+1}{n-m+1})$

As we can see, each term is at most $(\frac{n-k}{n})$.

Therefore, their product is at most $(\frac{n-k}{n})^m$.

Answer to Question 3(b)

We know that,
P(”don’t know” as min-hash value for this column) is at most $(\frac{n-k}{n})^m$
To prove: $-$
Probability of “don’t know” to be at most e^{-10} given $n >> m, k$

$(\frac{n-k}{n})^m <= e^{-10}$
$(1 - \frac{k}{n})^m <= e^{-10}$
$((1 - \frac{k}{n})^{\frac{n}{k}})^{\frac{mk}{n}} <= e^{-10}$
For large x, $(1 - \frac{1}{x})^x \approx \frac{1}{e}$
$(\frac{1}{e})^{\frac{mk}{n}} <= e^{-10}$
$e^{\frac{-mk}{n}} <= e^{-10}$
$\frac{-mk}{n} <= -10$
$\frac{mk}{n} >= 10$
$k >= \frac{10n}{m}$

So, $k = \frac{10n}{m}$ is a lower bound.

Answer to Question 3(c)

Consider 2 columns as follows,
S1 = [1, 0, 1, 0], S2 = [1, 1, 1, 0]

Jaccard Similarity is given by,
$= \frac{S1 \cap S2}{S1 \cup S2}$
$= \frac{2}{3}$

We can take random permutations,

Random Permutation	S1	S2
1 2 3 4	1	1
4 1 2 3	2	1
2 3 4 1	2	2
3 4 1 2	1	1

Probability of S1 and S2 have the same minhash values is $\frac{3}{4}$.