CS550: Massive Data Mining and Learning Spring 2020
Problem Set 1
Due 11:59pm Thursday, March 5, 2020

Only one late period is allowed for this homework (11:59pm Friday 3/6)

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):
Aayush Mandhyan

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

I acknowledge and accept the Honor Code.

(Signed)______________S.J____

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

Answer to Questions 1

1.) See the corresponding code file Q1.ipynb

2.)
In the first Map job, pairs of users who are already friends or have mutual friends are generated.
If they are already friends, a tuple ((user, friend), - 9999999999) is generated
and if they have mutual friends, a tuple ((user, friend), 1) is generated.

During the first Reduce job, the number of mutual friends are counted for each pair of users.
The keys are the pairs of users.
Since the value for pairs of users who are already friends to be – 9999999999, the sum for these pairs is always negative.
The reducer finds the sum of the values, if the sum of the values is less than 0 then the users are already friends, so we do not recommend these pairs of users to each other.

If the sum of the values is greater than 0 then we output the tuple ((user1, user2), number of mutual friends).
During the second Map job, ((user1, user2), number of mutual friends) is converted
to (user1, (user2, number of mutual friends)).
During the second Reduce job, groupByKey() is used where user1 is the key. During the 3rd Map job, the recommendation of at most 10 friends is mapped to each user.

3.) Recommendations for the 10 users:
a. 924: 439,2409,6995,11860,15416,43748,45881
b. 8941: 8943,8944,8940
c. 8942: 8939,8940,8943,8944
d. 9019: 9022,317,9023
e. 9020: 9021,9016,9017,9022,317,9023
f. 9021: 9020,9016,9017,9022,317,9023
g. 9022: 9019,9020,9021,317,9016,9017,9023
h. 9990: 13134,13478,13877,34299,34485,34642,37941
i. 9992: 9987,9989,35667,9991
j. 9993: 9991,13134,13478,13877,34299,34485,34642,37941

Answer to Questions 2(a)

For e.g. if A and B are independent of each other, then conf(A -> B) will become Pr(B).
In this case if the support for B is high, the conf(A -> B) will become a valid rule. So conf(A->B) becomes entirely dependent on B.
This is not the case when calculating lift or conviction because they both consider occurrences of both entities in their respective formulae.

Answer to Questions 2(b)

#Lift is symmetrical.

Proof: -
$lift(A{\rightarrow}B){=}conf(A{\rightarrow}B)S(B)$
=$P(B|A)P(B)$
=$P(A \cap B)P(B).P(A)$

Similarly,
$lift(B{\rightarrow}A){=}conf(B{\rightarrow}A)S(A)$
=$P(A|B)P(A)$
=$P(A \cap B)P(A).P(B)$

#Confidence and Conviction are not symmetrical.
Suppose we have 5 baskets.
B1 = (A,D), B2 = (B,E), B3 = (A,B), B4 = (C,E), B5 = (B,D) having 2 items each.
Here consider S(A) = 2/5, S(B) = 3/5, P(A ∩ B) = 1/5
confidence (A -> B) = (1/5) / (2/5) = 1/2
AND

confidence (B -> A) = (1/5) / (3/5) = 1/3
Hence, Confidence is not Symmetrical.

conviction (A -> B) = (1 – S(B)) / (1 – conf (A->B))
= (1 – 3/5) / (1 – 1/2) = 4/5
Similarly,
conviction (B -> A) = (1 – S(A)) / (1 – conf (B->A))
= (1 – 2/5) / (1 – 1/3) = 9/10

Hence, Conviction is not Symmetrical.

Answer to Questions 2(c)

Consider a rule A → B, where B occurs every time A occurs, and B can also occur even if A doesn't.
Then,

$conf(A \rightarrow B) = Pr(B|A) / Pr(A) = 1$

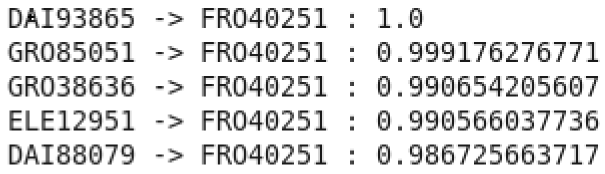
$conv(A \rightarrow B) = (1 - Pr(B)) / (1 - conf(A \rightarrow B)) = \infty$

lift(B->A) depends on the value of Pr(B) and may differ as B might occur in
baskets which do not have A.
Example:
If we have baskets AB, CD, CD, EF, then we have S(B)=1/4, S(D)=1/2, Pr(B|A)=1,
Pr(D|C)=1, then $lift(A \rightarrow B)$ = 1/0.25 = 4
& $lift(C \rightarrow D)$ = 1/0.5 = 2.

Here both rules have 100% rules, but they have different lift scores.
Therefore, Confidence and Conviction help identify the best rules when compared to Lift.

Answer to Questions 2(d)

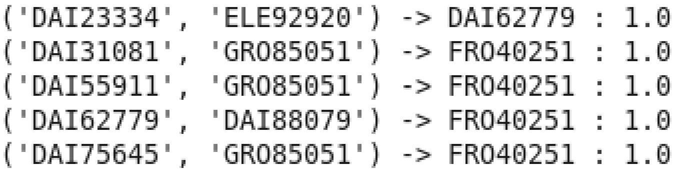
Top 5 pairs by confidence



```
DAI93865 -> FRO40251 : 1.0
GRO85051 -> FRO40251 : 0.999176276771
GRO38636 -> FRO40251 : 0.990654205607
ELE12951 -> FRO40251 : 0.990566037736
DAI88079 -> FRO40251 : 0.986725663717
```

Answer to Questions 2(e)

Top 5 triplets by 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 Questions 3(a)

Number of columns with m 1's out of n columns can be given as $\binom{n}{m}$.
Out of $\binom{n}{m}$ columns, the number of columns that have no 1 in one of the k selected rows is $\binom{n-k}{m}$.
The probability of no 1 in the chosen k rows is

$$P(\text{no } 1) = \frac{\binom{n-k}{m}}{\binom{n}{m}}$$

$$= \frac{\frac{(n-k)!}{m!(n-k-m)!}}{\frac{n!}{m!(n-m)!}}$$

$$= \frac{(n-k)!(n-m)!}{(n-k-m)!n!}$$

$$= \frac{(n-k)(n-k-1)\ldots(n-k-m+1)}{n(n-1)\ldots(n-m+1)}$$

Each of the m factors is at most $\left(\frac{n-k}{n}\right)$

And therefore, their product is at most $\left(\frac{n-k}{n}\right)^{m}$

Answer to Questions 3(b)

Since we can use $\left(\frac{n-k}{n}\right)^{m}$ as exact probability, we want to find k such that

$$= \left(\frac{n-k}{n}\right)^{m} \leq \; e^{-10}$$

Multiplying and dividing the exponent by $\frac{n}{k}$, we get

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

Since we assume n is much larger than k, we can approximate$\left(1-\frac{k}{n}\right)^{\frac{n}{k}}$ by *1/e*

We get, $e^{\frac{-mk}{n}} \leq e^{-10}$

Implies, $\frac{mk}{n} \geq 10$

Therefore, $k \geq \frac{10n}{m}$

Answer to Questions 3(c)

The two (sets) are $\begin{bmatrix}0\\1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\\1\end{bmatrix}$.
The Jaccard similarity is 0.5.
If the cycle starts at first or second row, we will get same minhash value but if the cycle starts at the last row we will get different minhash values. Hence the probability of minhash values agreeing in 2/3 if we continue in cyclic manner.