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):
Yanhan Zhang

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

I acknowledge and accept the Honor Code.

(Signed)____YL_____

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

Answer to Questions 1

(i) The source code is attached.

(ii) In the class Map, it generates two kind of relations. The pair of the user and the user's friend is marked 1. The pair of two friends of the user is marked 2.
In the class Reduce, first it counts the amount of each user's mutual friends. Then it sorts the value of each mutual friend and output the 10 mutual friends who have the largest number of mutual friends in common with the user.

(iii) The Recommendations for the users with following user IDs:

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 Questions 2(a)

As the definition of confidence, conf(A→B) = Pr(B|A). While A and B are independent, Pr(B|A) = Pr(B). At this time, if Pr(B) is very high, the value of conf(A→B) will still be high even there is not a real association between A and B. Lift and conviction do not suffer from this drawback because they put Support(B) in the computation as a represent of Pr(B).

Answer to Questions 2(b)

Confidence: assume there are two baskets {A, B} and {A, C}. Then $\text{conf}(A\rightarrow B) = \frac{Support(A\cup B)}{Support(A)} = \frac{1}{2}$, $\text{conf}(B\rightarrow A) = \frac{Support(A\cup B)}{Support(B)} = 1$. Since conf(A→B) ≠ conf(B→A), confidence is not symmetrical.

Lift: $\text{lift}(A\rightarrow B) = \frac{conf(A\rightarrow B)}{S(B)} = \frac{S(A\cup B)}{S(A)} \times \frac{1}{S(B)} = \frac{S(A\cup B)}{S(A)S(B)}$, $\text{lift}(B\rightarrow A) = \frac{conf(B\rightarrow A)}{S(A)} = \frac{S(A\cup B)}{S(B)} \times \frac{1}{S(A)} = \frac{S(A\cup B)}{S(A)S(B)}$. Since lift(A→B) = lift(B→A), lift is symmetrical.

Conviction: assume there are two baskets {A, B} and {A, C}. Then $\text{conf}(A\rightarrow B) = \frac{Support(A\cup B)}{Support(A)} = \frac{1}{2}$, $\text{conf}(B\rightarrow A) = \frac{Support(A\cup B)}{Support(B)} = 1$, $S(A) = \frac{Support(A)}{N} = 1$, $S(B) = \frac{Support(B)}{N} = \frac{1}{2}$. We have $\text{conv}(A\rightarrow B) = \frac{1-S(B)}{1-conf(A\rightarrow B)} = 1$, $\text{conv}(B\rightarrow A) = \frac{1-S(A)}{1-conf(B\rightarrow A)} = \frac{0}{0}$ which is infinity. Since conv(A→B) ≠ conv(B→A), conviction is not symmetrical.

Answer to Questions 2(c)

Confidence and conviction are desirable. If every time when A occurs, B always occurs. The value of conf(A→B) is 1 which is maximal. The value of conv(A→B) is infinity which is maximal. So, both confidence and conviction are desirable.

Answer to Questions 2(d)

The top 5 rules:

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)

The top 5 rules:

(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)

The probability of getting "don't know" as the min-hash value when random choose k of n rows equals to $\frac{\binom{n-m}{k}}{\binom{n}{k}} = \frac{\frac{(n-m)!}{(n-m-k)!k!}}{\frac{n!}{(n-k)!k!}} = \frac{(n-k-m+1)\times(n-k-m+2)\times\cdots\times(n-k)}{(n-m+1)\times(n-m+2)\times\cdots\times n} = \frac{n-k-m+1}{n-m+1} \times \frac{n-k-m+2}{n-m+2} \times \cdots \times \frac{n-k}{n}$. It is obvious that each item of this formula is no larger than $\frac{n-k}{n}$. When m = 0, the value of the formula is $(\frac{n-k}{n})^m$, which is maximal. So, the probability is at most $(\frac{n-k}{n})^m$.

Answer to Questions 3(b)

According to the question, we want: $\left(\frac{n-k}{n}\right)^m \leq e^{-10}$.We have $\left(\frac{n-k}{n}\right)^m = \left(1-\frac{k}{n}\right)^m = \left(\left(1-\frac{k}{n}\right)^{\frac{n}{k}}\right)^{\frac{mk}{n}}$. Since n is much larger than k, we can approximate $\left(1-\frac{k}{n}\right)^{\frac{n}{k}}$ to $\frac{1}{e}$. Then we have $(e^{-1})^{\frac{mk}{n}} \leq e^{-10}$, which can be represented as $\frac{mk}{n} \geq 10$. Then we have $k \geq \frac{10n}{m}$. So, the smallest value of k is $\frac{10n}{m}$.

Answer to Questions 3(c)

The two columns are $[1,0,1]^T$ and $[1,0,0]^T$. The Jaccard similarity is $\frac{1}{2}$. The random cyclic permutations can be (1, 2, 3), (3, 1, 2) and (2, 3, 1) and the corresponding min-hash values are (1, 1), (2, 3) and (1, 2). So, the probability is $\frac{1}{3}$.