CSE 547: Machine Learning for Big Data
Homework 1

Academic Integrity We take academic integrity extremely seriously. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom they interacted with.

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

- *Hwai-Jin Peng*

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

- CSE 547 Lecture Slide (Tim Althoff)
- Mining of Massive Datasets (Leskovec-Rajaraman-Ullman): Chapter 2
- Mining of Massive Datasets (Leskovec-Rajaraman-Ullman): Chapter 6
- Introduction to Data Mining (Tan-Steinbach-Karpatne-Kumar): Chapter 6

I acknowledge and accept the Academic Integrity clause.

(Signed) ***Cheng-Yen Yang***

Answer to Question 1

(a)

1. First we read the data from the text file followed by a map function to construct pairs with different x which represent different type of connections between the users. $((u_1, u_2), x)$ with $x = 0$ indicates that these two users share direct connection (first-degree friend) while $x = 1$ indicates that these two users share a mutual friend (second-degree friend).

2. Then we use GroupByKey and filter function to remove those $((u_1, u_2), 1)$ pairs if $((u_1, u_2), 0)$ existed. By doing so we prevent from recommending users whom are already friends.

3. Then we use ReduceByKey function to sum the number of mutual friend for each pair where tuple became $((u_1, u_2), n)$.

4. In order to give recommendations in descending order of the number of mutual friends, we use another simple map function to make the tuple into $(u_1, (u_2, n))$ and $(u_2, (u_1, n))$.

5. Last, we use GroupByKey function again to group the recommendations of the same user together then apply sort in descending order of number of mutual friends then user IDs in numerically ascending order.

(b)

User IDs	Recommendations
924	439, 2409, 6995, 11860, 15416, 43478, 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)

The drawback of ignoring $Pr(B)$ may resulted in misinterpretation of the association rules. For example, if the occurrence of B is totally unrelated to the occurrence of A, said $Pr(A)$ and $Pr(B)$ are two independent events, we'll have:

$$conf(A \to B) = Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{Pr(A) \times Pr(B)}{Pr(A)} = Pr(B). \tag{1}$$

And if B have a relatively high support, the value of $conf(A \to B)$ will be entirely decide by $Pr(B)$ even though these two events are totally unrelated.

Lift take $Pr(B)$ into consideration in the sense of measuring how much more "A and B occur together" than "what would be expected if A and B were statistically independent" while *conviction* compares the "probability that A appears without B if they were independent" with the "actual frequency of the appearance of A without B". Besides the definition, we can also observe that both *lift* and *conviction* take $Pr(B)$ into account in their equations:

$$lift(A \to B) = \frac{conf(A \to B)}{support(B)/N} \quad \text{and} \quad conv(A \to B) = \frac{1 - support(B)/N}{1 - conf(A \to B)}. \tag{2}$$

Answer to Question 2(b)

Confidence is asymmetrical. By the definition we have:

$$conf(A \to B) = Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} \quad \text{and} \quad conf(B \to A) = Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}, \tag{3}$$

we can give a counter example by having different $Pr(A)$ and $Pr(B)$ like:

$$Basket_1 = \{A, B\},\ Basket_2 = \{A, C\} \tag{4}$$

we have $conf(A \to B) = \frac{1}{2}$ and $conf(B \to A) = 1$.

Confidence is symmetrical. By the definition and following derivation we have:

$$lift(A \to B) = \frac{conf(A \to B)}{Pr(B)} = \frac{Pr(B|A)}{Pr(B)} = \frac{Pr(A \cap B)}{Pr(A) \times Pr(B)} \tag{5}$$

and

$$lift(B \to A) = \frac{conf(B \to A)}{Pr(A)} = \frac{Pr(A|B)}{Pr(A)} = \frac{Pr(A \cap B)}{Pr(A) \times Pr(B)}. \tag{6}$$

Conviction is asymmetrical. And we can use the counter example:

$$Basket_1 = \{A, B\},\ Basket_2 = \{A, C\},\ Basket_3 = \{C\}, \tag{7}$$

we have

$$conv(A \to B) = \frac{1 - conf(A \to B)}{1 - Pr(B)} = \frac{1 - \frac{1}{2}}{1 - \frac{2}{3}} = \frac{3}{2} \tag{8}$$

and

$$conv(B \to A) = \frac{1 - conf(B \to A)}{1 - Pr(A)} = \frac{1 - 1}{1 - \frac{2}{3}} = 0. \tag{9}$$

Answer to Question 2(c)

Confidence and *conviction* are desirable while *lift* is not. A measure is desirable if it reaches its maximum achievable value for all perfect implications, if Y occurs every single time X occurs, we can have the values of each measure being:

$$conf(X \to Y) = 1 \quad \text{and} \quad conv(X \to Y) = \infty, \tag{10}$$

and $lift(X \to Y)$ being undetermined (depending on $Pr(Y)$).

$$Basket_1 = \{A, B\},\ Basket_2 = \{A, B\},\ Basket_3 = \{C, D\}, \tag{11}$$

Given the above example, we observe that B occurs every single time A and D occurs every single time C occurs. However, we have that

$$lift(A \to B) = \frac{conf(A \to B)}{Pr(B)} = \frac{1}{1/2} = 2 \quad \text{and} \quad lift(C \to D) = \frac{conf(C \to D)}{Pr(D)} = \frac{1}{1/3} = 3, \tag{12}$$

although these two rules both hold 100% of the time, they have distinct *lift* values.

Answer to Question 2(d)

Rules $X \rightarrow Y$	Confidence
DAI93865 $\rightarrow$ FRO40251	1.0
GRO85051 $\rightarrow$ FRO40251	0.999176276771005
GRO38636 $\rightarrow$ FRO40251	0.9906542056074766
ELE12951 $\rightarrow$ FRO40251	0.9905660377358491
DAI88079 $\rightarrow$ FRO40251	0.9867256637168141

Answer to Question 2(e)

Rules $(X, Y) \rightarrow Z$	Confidence
(DAI23334, ELE92920) $\rightarrow$ DAI62779	1.0
(DAI31081, GRO85051) $\rightarrow$ FRO40251	1.0
(DAI55911, GRO85051) $\rightarrow$ FRO40251	1.0
(DAI62779, DAI88079) $\rightarrow$ FRO40251	1.0
(DAI75645, GRO85051) $\rightarrow$ FRO40251	1.0

Answer to Question 3(a)

Given that a column has m ones and $(n-m)$ zeros, the total number of combinations of columns with m ones in n rows is $\binom{n}{m}$ and the total number of combinations of getting "don't know" as the min-hash value is $\binom{n-k}{m}$ as columns have no ones in any of the k rows.
Therefore we have the probability:

$$Pr(X) = \frac{\binom{n-k}{m}}{\binom{n}{m}} = \frac{\frac{(n-k)!}{m!(n-k-m)!}}{\frac{n!}{m!(n-m)!}} = \frac{(n-m)!(n-k)!}{n!(n-m-k)!} = \frac{(n-m)!}{n!} \times \frac{(n-k)!}{(n-m-k)!}, \tag{13}$$

we can deduce $\frac{(n-m)!}{n!}$ and $\frac{(n-k)!}{(n-m-k)!}$ into:

$$Pr(X) = \frac{(n-k) \times (n-k+1) \times \cdots \times (n-k-m+1)}{n \times (n-1) \times \cdots \times (n-m+1)} \leq \Big(\frac{n-k}{n}\Big)^m, \tag{14}$$

as there are m pairs and each numerical value of them will be at most $\frac{n-k}{n}$.

Answer to Question 3(b)

From part(a):

$$Pr(X) \leq \Big(\frac{n-k}{n}\Big)^m = \Big(1-\frac{k}{n}\Big)^m, \tag{15}$$

from the definition of e we known that:

$$e = \lim_{n\to\infty} \big(1+\frac{1}{n}\big)^n, \tag{16}$$

so given $n >> m$ and $n >> k$ we have:

$$Pr(X) \leq \Big[\Big(1-\frac{k}{n}\Big)^{\frac{-n}{k}}\Big]^{\frac{-mk}{n}} = e^{\frac{-mk}{n}}. \tag{17}$$

If we want $Pr(X)$ to be at most e^{-10},

$$\frac{-mk}{n} \leq -10 \Longleftrightarrow \frac{10n}{m} \leq k, \tag{18}$$

the approximation of smallest value of k will be $\frac{10n}{m}$.

Answer to Question 3(c)

Given the input matrix S_1 and S_2 with Jaccard similarity being $\frac{1}{3}$:

S_1	S_2
0	0
1	0
0	1
1	1

and the cyclic permutations π_1,π_2, π_3 and π_4:

π_1	π_2	π_3	π_4
1	4	3	2
2	1	4	3
3	2	1	4
4	3	2	1

we can construct the signature matrix using the permutations and obtain that the probability that a random cyclic permutation yields the same min-hash value for both S_1 and S_2 is $\frac{1}{4}$:

	S_1	S_2
π_1	2	3
π_2	1	2
π_3	2	1
π_4	1	1