CS550: Massive Data Mining and Learning Spring 2019
Problem Set 1
Due 11:59pm Saturday, March 2, 2019

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

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):

None

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

Hypergeometric distribution: https://en.wikipedia.org/wiki/Hypergeometric_distribution

I acknowledge and accept the Honor Code.

(Signed)_Haoyang Zhang____

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

Answer to Questions 1

The basic idea of this algorithm is that: for each line, say `user, friendList`, all pairs in the `friendList` has a common friend `user`. Therefore, we can use each line to creat a number of key-value pair, where the key is the pair of the `friendList`, and the value is `1` to indicate we saw this pair for `1` time. Then we can use the idea of “word count” to get the number of times of each pair we saw. Notice that for each friend pairs, we should add one more key-value pair `(pair, -inf)` to avoid them presented in the result.

The recommendations for some users are shown below:

```
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)

Notice that $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P}(B)\mathrm{P}(\mathrm{A}|\mathrm{B})}{\mathrm{P}(\mathrm{A})}$. If $\mathrm{P}(\mathrm{B})$ is large enough, it still could be a large number. Namely, if everyone buys B, we will definitely have $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) = 1$. However, it does not make much sense.

Notice that $\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})}{\mathrm{P}(\mathrm{B})}$, which indicates how much more confidence we have when we know that person has bought A. The only case $\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B})$ is a large number is that B is strongly related to A. Specifically, the presence of A is a strong positive signal of the presence of B.

Notice that $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) = \frac{\mathrm{P}(\bar{B})}{\mathrm{P}(\bar{B}|\mathrm{A})}$, which indicates how confident we will be if we know that person has bought A. The only case $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B})$ is a large number is that B is strongly related to A. Specifically, the presence of A is a strong negative signal of the absence of B.

Answer to Questions 2(b)

Since $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A})$, Say there are following case:

	A	not A
B	30	20
not B	0	50

Here we have $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) = 1$, $\mathrm{conf}(\mathrm{B} \rightarrow \mathrm{A}) = 0.6$. Therefore, $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) \neq \mathrm{conf}(\mathrm{B} \rightarrow \mathrm{A})$.

$$\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})}{\mathrm{P}(\mathrm{B})} = \frac{P(AB)}{P(A)P(B)} = \frac{\mathrm{P}(\mathrm{A}|\mathrm{B})}{\mathrm{P}(\mathrm{A})} = \mathrm{lift}(\mathrm{B} \rightarrow \mathrm{A})$$

Notice that in previous example, $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) = +\infty$, $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) = 1.75$. Therefore, $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) \neq \mathrm{conv}(\mathrm{B} \rightarrow \mathrm{A})$

Answer to Questions 2(c)

Notice that $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) \in [0,1]$. When $\mathrm{P}(\mathrm{B}|\mathrm{A}) = 1$, $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B}) = 1$. Therefore, $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B})$ is desirable.

$\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B}) \in [0, +\infty)$. When $\mathrm{P}(\mathrm{B}|\mathrm{A}) = 1$, $\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B}) = \frac{1}{\mathrm{P}(\mathrm{B})}$. Assuming $\mathrm{P}(\mathrm{A}) \neq 0, \mathrm{P}(\mathrm{B}) \neq 0$, $\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B}) < +\infty$. Therefore, $\mathrm{lift}(\mathrm{A} \rightarrow \mathrm{B})$ is not desirable.

$\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) \in [0, +\infty)$. When $\mathrm{P}(\mathrm{B}|\mathrm{A}) = 1$, $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) = \frac{1-\mathrm{P}(\mathrm{B})}{0}$. Assuming $\mathrm{P}(\mathrm{A}) \neq 1, \mathrm{P}(\mathrm{B}) \neq 1$, $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B}) = +\infty$. Therefore, (if $\mathrm{P}(\mathrm{B}) \neq 1$,) $\mathrm{conv}(\mathrm{A} \rightarrow \mathrm{B})$ is desirable.

Answer to Questions 2(d)

```
DAI93865 -> FRO40251: 1.000
GRO85051 -> FRO40251: 0.999
GRO38636 -> FRO40251: 0.991
ELE12951 -> FRO40251: 0.991
DAI88079 -> FRO40251: 0.987
```

Answer to Questions 2(e)

```
DAI23334 + ELE92920 -> DAI62779: 1.000
DAI31081 + GRO85051 -> FRO40251: 1.000
DAI55911 + GRO85051 -> FRO40251: 1.000
DAI62779 + DAI88079 -> FRO40251: 1.000
DAI75645 + GRO85051 -> FRO40251: 1.000
```

Answer to Questions 3(a)

Notice that the distribution of number of 1’s we get, say X, when we choose k rows is hypergeometric distribution. Specifically, $\mathrm{X} \sim \mathrm{Hypergeometric}(\mathrm{n}, \mathrm{m}, \mathrm{k})$.

$$\mathrm{P}(\mathrm{X}=0)=\frac{C_m^0 C_{n-m}^k}{\mathrm{C}_{\mathrm{n}}^{\mathrm{k}}}=\frac{\frac{(\mathrm{n}-\mathrm{m})!}{(\mathrm{n}-\mathrm{m}-\mathrm{k})!}}{\frac{\mathrm{n}!}{(\mathrm{n}-\mathrm{k})!}}=\frac{(n-k)(n-k-1)\ldots(n-k-m+1)}{n(n-1)\ldots(n-m+1)} \leq \left(\frac{n-k}{n}\right)^m$$

Answer to Questions 3(b)

$$\mathrm{p} = \left(\frac{\mathrm{n}-\mathrm{k}}{\mathrm{n}}\right)^{\mathrm{m}} = \left(1 - \frac{1}{\frac{n}{k}}\right)^{\frac{n}{k}\frac{m}{\frac{n}{k}}} \approx e^{-\frac{m}{\frac{n}{k}}} \leq e^{-10}$$

$$\frac{\mathrm{m}}{\frac{\mathrm{n}}{\mathrm{k}}} \geq 10$$

$$\mathrm{k} \geq 10\frac{\mathrm{n}}{\mathrm{m}}$$

Answer to Questions 3(c)

Here is the transposition of that matrix: (Using row vectors, instead of column vectors, to save room.)

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0
B	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	1	0	1

Here we have $\mathrm{sim}(A,B) = \frac{1}{10}$. But if we choose $r = 1, 2, 3, \dots, 11$, we will have $h(A) = h(B)$. Namely, $P\big(h(A) = h(B)\big) = \frac{11}{20} \gg \mathrm{sim}(A, B)$