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):

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

I acknowledge and accept the Honor Code.

(Signed)_____RK_____

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

Answer to Questions 1

There are two python files: one is mapper.py and the other is reducer.py.

Mapper.py

We find out all pairs of possible friends in this approach. In this approach an indicator is maintained and the direct friends are indicated as 1 and indirect friends are indicated as 0. The one's indicated as 0 are basically formed by the combination of friend's friends or mutual friends as they say.

Reducer.py

At this point, we group all the key and value pair. Basically, we count the number of times a pair have mutual friends. Based on the indicator value we will decide if the people are already friends or not. We will count only if there is no connection between them. In this way we find the mutual friends. Dictionaries are used to store the mutual friends.

Recommendations

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)

The fact that confidence ignores probability of B is a disadvantage as some times in the events of two events being independent P (B|A) is same as P(B) which is high and the rule A->B is considered to be legit. But lift and conviction takes the probability of B into consideration.

Answer to Questions 2(b)

LHS

$$\boldsymbol{lift}\ (\ \boldsymbol{A \rightarrow B}) = \frac{\boldsymbol{conf}\ (\boldsymbol{A \rightarrow B})}{\boldsymbol{S(B)}}$$

$$= \frac{\boldsymbol{P(B|A) * N}}{\boldsymbol{Support}(\ \boldsymbol{B}\ )}$$

$$= \frac{\boldsymbol{P(A \cap B) * N}}{\boldsymbol{P(A) *\ Support(B)}}$$

$$\frac{\boldsymbol{P(A \cap B) * N^2}}{\boldsymbol{Support(A) * Support(B)}}$$

RHS

$$\boldsymbol{lift}\ (\ \boldsymbol{B \rightarrow A}) = \frac{\boldsymbol{conf}\ (\boldsymbol{B \rightarrow A})}{\boldsymbol{S(A)}}$$

$$= \frac{\boldsymbol{P(A|B) * N}}{\boldsymbol{Support}(\ \boldsymbol{A})}$$

$$= \frac{\boldsymbol{P(A \cap B) * N}}{\boldsymbol{P(B) *\ Support(A)}}$$

$$\frac{\boldsymbol{P(A \cap B) * N^2}}{\boldsymbol{Support(A) * Support(B)}}$$

As LHS = RHS , lift is symmetric.

$$\boldsymbol{conf}\ (\boldsymbol{A \rightarrow B})\ \ \boldsymbol{is}\ \boldsymbol{P(B|A)}\ \boldsymbol{and}\ \boldsymbol{conf}\ (\boldsymbol{B \rightarrow A})\ \boldsymbol{is}\ \boldsymbol{P}\ (\boldsymbol{A|B})$$

They need not be the same necessarily. So, confidence is not symmetric.

As conviction is based on confidence, we can deduce that conviction is not symmetric.

$$\boldsymbol{conv\,(\,A \rightarrow B) = \frac{1 - S(B)}{1 - conf(A \rightarrow B)}}$$

$$\frac{\mathbf{1} - \frac{\boldsymbol{Support\,(\,B)}}{\boldsymbol{N}}}{\mathbf{1} - \boldsymbol{P(B|A)}}$$

$$\boldsymbol{conv\,(\,B \rightarrow A) = \frac{1 - S(A)}{1 - conf(B \rightarrow A)}}$$

$$\frac{\mathbf{1} - \frac{\boldsymbol{Support\,(\,A)}}{\boldsymbol{N}}}{\mathbf{1} - \boldsymbol{P(A|B)}}$$

Clearly both the values are not same. So , conviction is not symmetric.

Answer to Questions 2(c)

Lift is not desirable and conviction and confidence are desirable.

In confidence,

We know that P(A $\cap \boldsymbol{B}) = \boldsymbol{P}\,(\boldsymbol{A})$

$$\boldsymbol{conf\,(\,A \rightarrow B) = P(B|A) = 1}$$

In conviction,

When we plug in the values,

$$\boldsymbol{conv\,(A \rightarrow B) = \infty}$$

As the lift value depends on the P(B), even though they are 100 % rules, they have different lift scores.

Example:

$$P(B|A) = 1$$

$$S(B) = 1/4$$

$$P(D|C) = 1$$

$$S(D) = 1/2$$

$$lift(A \rightarrow B) = 4$$

$$lift(B \rightarrow A) =$$

Answer to Questions 2(d)

Pairs	Confidence
DAI93865 FRO40251	1.0
GRO85051 FRO40251	0.999176276771005
GRO38636 FRO40251	0.9906542056074766
ELE12951 FRO40251	0.9905660377358491
DAI88079 FRO40251	0.9867256637168141
FRO92469 FRO40251	0.983510011778563

Answer to Questions 2(e)

DAI23334 ELE92920 DAI627792	1.0
DAI31081 GRO85051 FRO40251	1.0
DAI55911 GRO85051 FRO40251	1.0
DAI62779 DAI88079 FRO40251	1.0
DAI75645 GRO85051 FRO40251	1.0
GRO85051 SNA80324 FRO40251	1.0

Answer to Questions 3(a)

The number of columns with m ones out of n is ${}_{m}^{n}C$ **.**
To select columns with 1 in k rows the number of combinations is ${}_{m}^{n-k}C$
The probability of 1 in k rows is ${}_{m}^{n-k}C / {}_{m}^{n}C$
On simplification, we have:

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

On further simplification,

We have,

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

Each of these factors do not exceed $\frac{n-k}{n}$ **implying that the product will not exceed** $\frac{n-k^m}{n}$

Answer to Questions 3(b)

It is desired that $\frac{n-k^m}{n} \leq e^{-10}$

$$1 - \frac{k^m}{n} \leq e^{-10}$$

On multiplication and division by n/k, we have:

$$1 - \frac{k^{n/k}}{n}^{mk/n} \leq e^{-10}$$

As k is negligible with respect to n:

We approximate $1 - \frac{k^{n/k}}{n}$ **by 1/e.**

Which gives us:

$$e^{-mk/n} \leq e^{-10}$$

This implies that:

$$-\frac{mk}{n} \leq -10$$

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

$k \geq 10n/m$ **which is the lower bound for k.**

Answer to Questions 3(c)
Example:

1	**0**
0	0
1	0
0	1
0	0
1	1

The Jaccard similarity for the above example is equal to ¼.
The signature matrix for the above example will be:

1	4
2	3
1	2
3	1
2	2
1	1

The similarity for signature matrix 2/6 that is equal to 1/3.
Which is not equal to Jaccard Similarity.