CS550: Massive Data Mining and Learning
Homework 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):

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

The book and slides https://towardsdatascience.com/association-rules-2-aa9a77241654

I acknowledge and accept the Honor Code.

(Signed) FYMA _____

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

Answer to Question 1

question_1.java
MapReduce Algorithm Recommends the top 10 friends based on the number of mutual friends
The Algorithm :

Map Procedure:
- Loop over all the users list:
- Pick the user A:
- Loop over the user A friends:
- Pair the user A to the already friend users and 0: Key A, Pair (friend1, 0)
- Loop over each friend of user A friends and Pair them together with 1
which means they have one mutual friend A: Key: friend1, Pair (friend2,1)

The resulted pairs will be shuffled and sorted, combined by the similar keys and sent to the produce

Produce Procedure:
- Get a key with all its pairs
- Loop over these pairs:
- If the pair with count 0, ignore it because it means the user and
suggested are already friends
- If the pair with count=1, if it was counted before increase the counter,
else add it with count 1
- Will end up for key A : (friend1,1), (friend2,4), (friend3,9), where
friend1 is suggested and count is how many mutual friends between them
- Sort the friends by their mutual friends count and pick the top 10

Output Examples:

```
924 439,2409,6995,11860,15416,43748,45881
8941 8938,8942,8946,8939,8943,8944,8945,8940
8942 8938,8939,8941,8945,8946,8940,8943,8944
9019 320,9018,9016,9017,9020,9021,9022,317,9023
9020 9021,320,9016,9017,9018,9019,9022,317,9023
9021 9020,320,9016,9017,9018,9019,9022,317,9023
9022 9019,9020,9021,317,320,9016,9017,9018,9023
9990 9987,9988,9989,9993,9994,35667,9991,9992,13134,13478
9992 9987,9989,9988,9990,9993,9994,35667,9991
9993 9990,9994,9987,9988,9989,9991,35667,9992,13134,13478
```

Answer to Question 2(a)

$$conf(A \rightarrow B) = P(B|A) = P(BandA)/P(A)$$

Confidence ignoring P(B) will give a misleading rule because the

$$conf(A \rightarrow B)$$

might be high because it happens that product A and B to occur together very often (high P(A and B)) or B occurs very often. For example if

$$P(milkandegg) = .15$$

$$P(milk) = .2$$

$$P(egg) = .9$$

$$conf(milk \rightarrow egg) = .75$$

but if P(egg)=.9 this means it appears in the basket very often and this rule is misleading where

$$conf(notmilk \rightarrow egg)$$

can be high as well Support(B) which is used in both lift and conviction it is using S(B) which is the probability of B divided by the number of baskets giving the real numbers.

Answer to Question 2(b)

Confidence : this measure isn't symmetric Having :

$$conf(A \rightarrow B) = P(B|A) = P(B \cap A)/P(A)$$

$$conf(B \rightarrow A) = P(A|B) = P(A \cap B)/P(B)$$

Unless P(A) = P(B) -which is very rarely to happen- these two values won't be equal. For the example above the probability of occurrence of eggs in the basket if the basket already contains milk isn't equal to the probability of occurrence of eggs in the basket if the basket already contains milk

Lift : this measure is symmetric, the proof mathematically Having :

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{P(B)} = \frac{P(B \cap A)}{P(A)} * \frac{1}{P(B)}$$

$$lift(B \rightarrow A) = \frac{conf(B \rightarrow A)}{P(A)} = \frac{P(A \cap B)}{P(B)} * \frac{1}{P(A)}$$

And these values are equal.

Conviction : this measure is not symmetric, the proof mathematically again Having :

$$Conviction(A \rightarrow B) = \frac{1 - P(B)}{1 - conf(A \rightarrow B)}$$

$$Conviction(B \rightarrow A) = \frac{1 - P(A)}{1 - conf(B \rightarrow A)}$$

And these values are not equal. A and B are independent if the conviction value is high and the conf is small.

Answer to Question 2(c)

Confidence is desired because it will be equal 1 if A and B always occurs together. Lift is not equal for every time A and B appears together because it depends on the probability of B as well, it will be maximal if the confidence is equal to the P(B). Conviction is desired, it's the inverse of the lift value and when it's equal to 1 it always mean that A doesn't relate to B.

Answer to Question 2(d)

$$[DAI93865, FRO40251] = 1.0$$
$$[FRO40251, DAI93865] = 0.053594434424117494$$
$$[GRO85051, FRO40251] = 0.999176276771005$$
$$[FRO40251, GRO85051] = 0.0$$
$$[GRO38636, FRO40251] = 0.9906542056074766$$
$$[FRO40251, GRO38636] = 0.027312548312290647$$
$$[ELE12951, FRO40251] = 0.9905660377358491$$
$$[FRO40251, ELE12951] = 0.027054882762174697$$
$$[DAI88079, FRO40251] = 0.9867256637168141$$
$$[FRO40251, DAI88079] = 0.11491883535171347$$

Answer to Question 2(e)

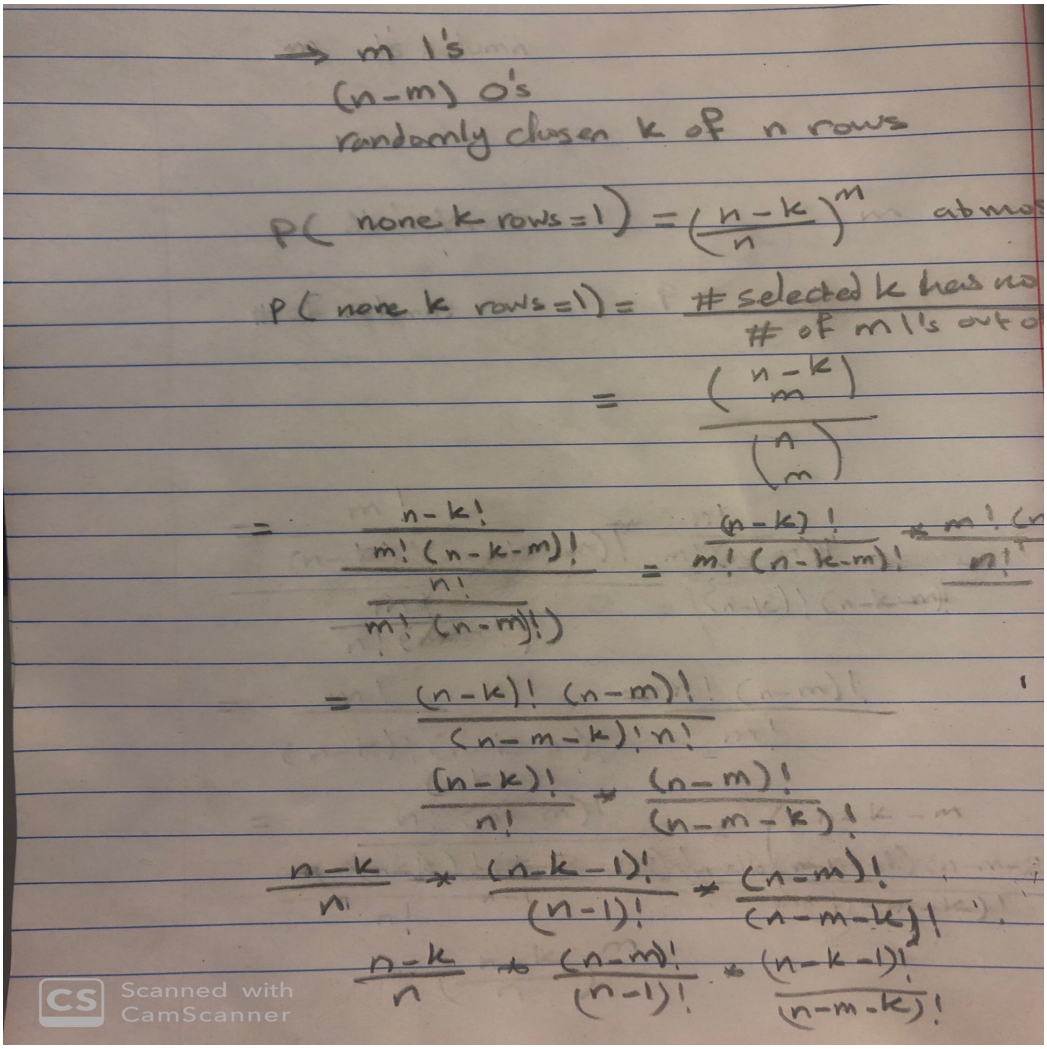
$$[DAI23334, ELE92920, DAI62779] = 1.0$$
$$[DAI23334, DAI62779, ELE92920] = 0.5238095238095238$$
$$[ELE92920, DAI62779, DAI23334] = 0.1630558722919042$$
$$[DAI31081, GRO85051, FRO40251] = 1.0$$
$$[DAI31081, FRO40251, GRO85051] = 0.36428571428571427$$
$$[GRO85051, FRO40251, DAI31081] = 0.08408903544929926$$
$$[DAI55911, GRO85051, FRO40251] = 1.0$$
$$[DAI55911, FRO40251, GRO85051] = 0.5732758620689655$$
$$[GRO85051, FRO40251, DAI55911] = 0.10964550700741962$$
$$[DAI62779, DAI88079, FRO40251] = 1.0$$
$$[DAI62779, FRO40251, DAI88079] = 0.10934579439252337$$
$$[DAI88079, FRO40251, DAI62779] = 0.2623318385650224$$
$$[DAI75645, GRO85051, FRO40251] = 1.0$$
$$[DAI75645, FRO40251, GRO85051] = 0.3149920255183413$$
$$[GRO85051, FRO40251, DAI75645] = 0.325638911788953$$

Answer to Question 3(a)

$P(nonekrows = 1) = (n - k/n)^m$
P(none k rows) = #k not 1/# of m 1's out of n

$$= (\binom{n-k}{m})/(\binom{n}{m})$$

$$= (n-k)/n * ((n-m)!/(n-1)!) * ((n-k-1)!/(n-m-k)!)$$



Answer to Question 3(b)

$$= ((n-k)/n)^m <= e^-10$$

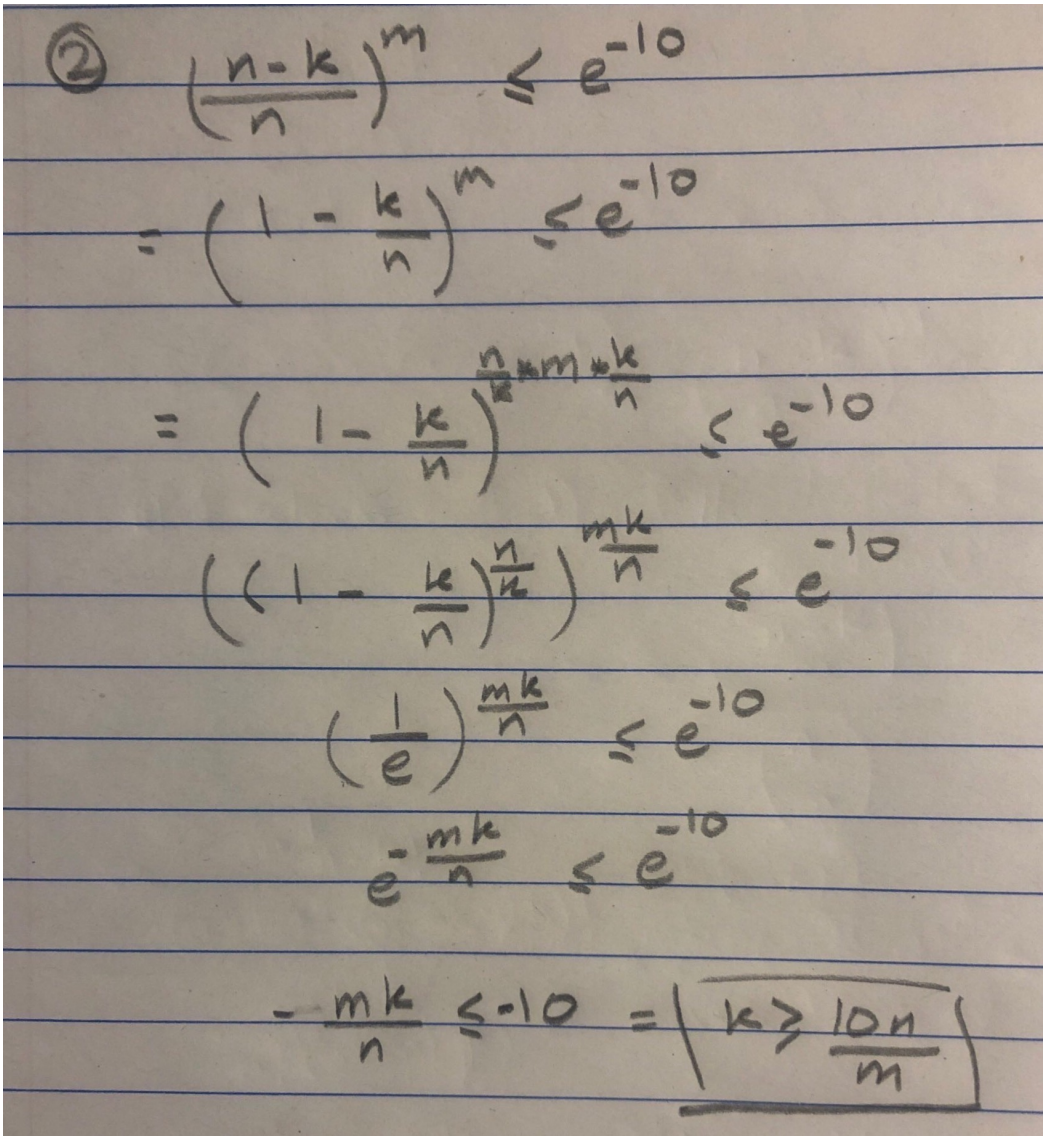
$$= (1-k/n)^m <= e^-10$$

$$(1-k/n)^{((n/k)*m*(k/n))} <= e^-10$$

$$(1/e)^{(mk/n)} <= e^-10$$

$$e^{(-mk/n)} <= e^-10$$

$$= k >= 10n/m$$



Answer to Question 3(c)