CS550: Massive Data Mining and Learning Spring 2023

Problem Set 1

Due 11:59pm Monday, February 20, 2023
Please see the homework file for late policy

Honor Code: Students may have discussions about the homework with peers. 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 with whom they have discussions about the homework. Directly using the code or solutions obtained from the web or from others is considered an honor code violation. We check all the submissions for plagiarism and take the honor code seriously, and we hope students to do the same.

Discussions (People with whom you discussed ideas used in your answers): ***Adithya Shrivastava***

On-line or hardcopy documents used as part of your answers: ***(LRU) Mining Massive Data Sets by J. Leskovec, A. Rajaraman, J. D. Ullman***

I acknowledge and accept the Honor Code.

(Signed)_____***Sanchit Thakur***_____

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

Answer to Questions 1(a)

Code for this question has been attached, titled '***question_1.ipynb***'.

Answer to Questions 1(b)

The algorithm implemented works as described below:

- The main primary function created is called '***search_connections***'. This function works by first separating user U and its list of friends by splitting using the parameter tab. We use a default value minimum to represent the connection between user U and corresponding friends. The list of friends is further split by using the comma parameter (since it is comma separated). Finally, this function returns to us the connected pairs as well as the common pairs of friends.
- Then we flatten the output of previous function by calling the flatMap function, followed by merging using the reduceByKey function. After this we filter only those pairs of users which have at least 1 friend in common.
- Finally, we use the map function to return the top 10 friend recommendations to each user U.

Answer to Questions 1(c)

After running the program, the recommendations for user IDs are listed 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)

Confidence does not take Pr(B) into account, as seen from the formula. This is a major drawback since it would lead to rules and associations which have high confidence but are ultimately misleading and therefore are not of much use. With respect to Market Analysis, if we take salt into consideration. Since it is a staple ingredient of cooking, it would be purchased by a large number of people, thus it would be in many baskets. This would mean that we would end up with high value of confidence even for those items which are not associated with salt. Thus, it would lead to an incorrect association.

We can see that lift and conviction do not suffer this same problem simply due to the fact that both of them take Support(B) into account. We can see this in the formula of both, which have been provided. Lift has S (B) in its denominator and S(B) = Support(B)/N.
Similarly, Conviction has conf(A → B) in its denominator and we know that conf(A → B) = Pr(B|A). Therefore, both lift and conviction avoid this problem by accounting for support(B) or Pr(B).

Answer to Questions 2(b)

CONFIDENCE: The first measure Confidence is not symmetrical.

From the definition of confidence, we know that:

$$\text{conf}(A \rightarrow B) = \Pr(B|A) = \Pr(B \cap A) / \Pr(A)$$

$$\text{Similarly, conf}(B \rightarrow A) = \Pr(A|B) = \Pr(A \cap B) / \Pr(B)$$

$\Pr(B \cap A) = \Pr(A \cap B)$ but $\Pr(A) \neq \Pr(B)$, thus confidence is not symmetrical.

For eg., if Pr(A) = 4/10, Pr(B) = 3/10, Pr(A ∩ B) = 2/10.

conf(A → B) = (2/10) / (4/10) = 2/4 = 0.5

conf(B → A) = (2/10) / (3/10) = 2/3 = 0.67

Thus, it is proved that confidence measure is not symmetrical.

LIFT: The next measure Lift is symmetrical

We know,

$lift(A \rightarrow B)$ = conf(A → B) / S(B) = (S(AB) / S(A)) / (S(B) / N) = (S(AB) * N) / (S(A) * S(B))

$lift(\mathrm{B} \rightarrow \mathrm{A})$ = conf(B → A) / S(A) = (S(BA) / S(B)) / (S(A) / N) = (S(BA) * N) / (S(B) * S(A))

Therefore, we can see that $lift(A \rightarrow B) = lift(\mathrm{B} \rightarrow \mathrm{A})$

Thus, it is proved that lift measure is symmetrical

CONVICTION: The last measure conviction is not symmetrical

$$\text{conv}(A \rightarrow B) = (1 - S(B)) / (1 - \text{conf}(A \rightarrow B))$$

$$\text{conv}(B \rightarrow A) = (1 - S(A)) / (1 - \text{conf}(B \rightarrow A))$$

As we can clearly see the two are not equal

For eg., continuing on from the example used for Confidence, we get

conv(A → B) = (1 – (3/10)) / (1 – (1/2)) = 1.4

conv(B → A) = (1 – (4/10)) / (1 – (2/3)) = 0.89

As we can see, the two are definitely not equal.

Thus, it is proved that conviction is not symmetrical.

Answer to Questions 2(c)

From the given measure, confidence and conviction are desirable and lift is not desirable.

- For Confidence, consider the scenario where B occurs every time A occurs. In such a scenario, Pr(A ∩ B) = Pr(A) which means we would get Pr(A) / Pr(A). So we can say that the confidence value is 1. Therefore, confidence can have a maximum value of 1.
- For Conviction, using the same scenario as above, if B occurs every time A also occurs, we would have conv(A → B) = 1, which means the denominator would become 1 – 1 = 0. Thus, the value of conviction becomes ∞, which is its maximum value.
- For Lift, we know that
  $lift(A \rightarrow B)$ = conf(A → B) / Pr(B)
  Since the lift depends on Pr(B), in the scenario above where B occurs every time A also occurs, in the case of lift, the value will depend on Pr(B). It decreases with increase in Pr(B), and thus might not be maximal in all cases. Hence, lift is not desirable for perfect implications.

Answer to Questions 2(d)

The top 5 pairs along with their confidence as obtained after running the program are shown below:

• DAI93865 → FRO40251; conf= 1.0
• GRO85051 → FRO40251; conf = 0.999176276771005
• GRO38636 → FRO40251; conf = 0.9906542056074766
• ELE12951 → FRO40251; conf = 0.9905660377358491
• DAI88079 → FRO40251; conf = 0.9867256637168141

Answer to Questions 2(e)

The top 5 triples along with their confidence as obtained after running the program are shown below:

- DAI23334, ELE92920 → DAI62779; conf = 1.0
- DAI31081, GRO85051 → FRO40251; conf = 1.0
- DAI55911, GRO85051 → FRO40251; conf = 1.0
- DAI62779, DAI88079 → FRO40251; conf = 1.0
- DAI75645, GRO85051 → FRO40251; conf = 1.0

Answer to Questions 3(a)

Given:
Column has m 1's and (n-m) 0's.
Therefore, number of ways of selecting k rows from n rows whose value is 0 becomes C^{n-m}_k
Also, total possible ways of selecting k rows from n rows is C^n_k
So hence we can say that the probability of getting "don't know" as minhash value for this column would be the same as probability of selecting k rows with value of 0.
This would be = $(C^{n-m}_k) / (C^n_k)$ = ((n-m)! / k!(n-m-k)!) / (n! / k! (n – k)!)
= ((n-m)! (n-k)!) / n! (n – m – k)! = ((n-k)/n)((n-k-1)/(n-1))…. ((n-k-m+1)/(n-m+1))

From here we can do a bit of approximations, as each term in the product is approximately equal to (n-k)/m. Also, as we can see, there are m terms in the product.
Therefore, we get the product = $((n-k)/n)^m$
Thus, it is proved that probability of getting "don't know" as minhash value is at most $((n-k)/n)^m$.

Answer to Questions 3(b)

Assumption: n and m are both very large
To find: smallest value of k at which probability is at most e^{-10}

As we know from the previous question, we have

$$((n-k) / n)^m \leq e^{-10} \qquad \ldots.(1)$$

We know that for large values of x

$$e^{-1} = (1 – 1/x)^x$$

Since n is much larger than k, we can say that x = n/k and thus we get

$$e^{-1} = (1 - 1/(n/k))^{n/k}$$

$$\Rightarrow e^{-10} = (1 – 1/(n/k))^{10n/k}$$

When we combine this with (1), we get:

$$((n-k) / n)^m \leq (1 – 1/(n/k))^{10n/k}$$

Since $(n-k)/n \leq 1$ we can say that $m \geq 10n/k$
Therefore, we get $k \geq 10n/m$
Hence, the smallest value of k is 10n/m

Answer to Questions 3(c)

The two columns that would be an example would be S1 = $[\ 0\ 1\ 0]^T$ and S2 = $[\ 0\ 1\ 1]^T$.

The Jaccard similarity of S1 and S2 = 0.5

Permutation	Minhash agree
1 2 3	Yes
1 3 2	No
2 3 1	Yes
2 1 3	Yes
3 2 1	No
3 1 2	No

As we can see, minhash values agree 3 times out of 6 which means 3/6 = 0.5, same as Jaccard similarity. But if we take cyclic permutations starting from last row of S1 and S2, minhash values would be different. In the case where the cycle starts at either of the one of the first two rows, the minshash values are the same, and in the case where the cycle starts at the last row, minhash values might differ. This means that the probability of the minhash values agreeing is 2/3 , when only cyclic permutations are allowed.