CS550: Massive Data Mining and Learning

Homework 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) AHM ______________________________

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

Answer to Question 1

1. **Source code:** It is attached in this zip file.

2. **Description:**

   We have 1 mapper and 3 reducers. Each one functions as follows:

   (a) First mapper maps the input line to $< (User, User1), 0 >$ if User and User1 are friends already, and to $< (User1, User2), 1 >$ if they are not friends directly.

   (b) First reducer receives all inputs with key as (User1, User2). We take a product of the complete output, and if it contained 0 at any point, then User1 and User2 are friends, and we don't output anything. If the product is not 0, then User1 and User2 are not friends with each other, and we output $< (User1, User2), sum(counts) >$.

(c) Next reducer receives input of the form $< (User1, User2), total\_counts >$. We output $< User1, (User2, total\_counts) >$.

(d) Last reducer receives all the links that have some mutual connection, along with total number of counts of those connection. This output is sorted and thrown to a file.

3. **Recommendation of particular users:**

   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 Question 2(a)

Confidence ignores $Pr(B)$. This is a drawback, as we can have cases where B occurs in all the baskets. Thus, having A in any basket doesn't imply or increase the chance of B in that basket, as the presence of B is independent of presence of A in a basket.

Lift doesn't suffer from this drawback, as

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

Thus, if B occurs in all the baskets, then the value of lift is lowered. On the other hand, if B occurs in only those baskets that have A, then the value of lift is maximum.

Conviction too doesn't suffer from this drawback, as

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

In this case, if B appears in all the baskets, then the value of conviction is 0. On the other hand, if B occurs in only those baskets that have A, then the value of conviction is

∞, agreeing with our intuition.

Answer to Question 2(b)

1. **Confidence:** Confidence is not symmetrical. As an example, consider the following market baskets: $\{A, B, C\}, \{A, C\}, \{A, B\}$. In this case, we have:

   $conf(A \rightarrow B) = Pr(B|A) = \frac{Pr(A,B)}{Pr(A)} = \frac{2}{3}$

   However, we also have:

   $conf(B \rightarrow A) = Pr(A|B) = \frac{Pr(A,B)}{Pr(B)} = 1$

   Hence, confidence is not symmetrical.

2. **Lift:** Lift is symmetrical. Proof:

   $lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{Pr(B|A)}{S(B)} = \frac{Pr(A,B)}{Pr(A)Pr(B)} \ldots$ (1)

   $lift(B \rightarrow A) = \frac{conf(B \rightarrow A)}{S(A)} = \frac{Pr(A|B)}{S(A)} = \frac{Pr(A,B)}{Pr(A)Pr(B)} \ldots$ (2)

   From (1) and (2), we have:

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

   $\therefore$ Lift is symmetrical.

3. **Conviction:** Conviction is not symmetrical. As an example, consider the following market baskets: $\{A, B, C\}, \{A, C\}, \{A, B\}, \{D\}$. In this case, we have:

   $conv(A \rightarrow B) = \frac{1-S(B)}{1-conf(A \rightarrow B)} = \frac{1-\frac{1}{2}}{1-\frac{2}{3}} = \frac{3}{2} \ldots$ (1)

   However, we also have:

   $conv(B \rightarrow A) = \frac{1-S(A)}{1-conf(B \rightarrow A)} = \frac{1-\frac{3}{4}}{1-1} = \infty \ldots$ (2)

   Hence, conviction is not symmetrical.

Answer to Question 2(c)

1. **Confidence:** For perfect implications, confidence has a maximum value of 1. Since it is a probability measure, for rules that occur together all the time, it's confidence is maximal and is 1. As an example consider the following market baskets:
   $\{A, B\}, \{A, B\}, \{C, D\}$
   Thus, $Pr(B|A) = 1$

2. **Lift:** Lift is not a desirable property, as for perfect implications, its value alters according to the denominator. As an example, consider the following market baskets:
   $\{A, B\}, \{A, B\}, \{C, D\}$
   Thus, $lift(A \rightarrow B) = \frac{1}{\frac{2}{3}} = \frac{3}{2}$
   However, $C \rightarrow D$ is also a perfect implication, but it's value of lift is even more.

$lift(C \rightarrow D) = \frac{1}{\frac{1}{3}} = 3$
$\therefore$ They have different lift scores.

3. **Conviction:** Conviction is a desirable property, as for perfect implications, it's value goes to infinity. As an example, consider the following market baskets:
$\{A, B\}, \{A, B\}, \{C, D\}$
Thus, $conv(A \rightarrow B) = \frac{1-\frac{2}{3}}{1-1} = \infty$

Answer to Question 2(d)

Top 10 pairs by confidence (only 5 needed):

The first column is the confidence score, the second column is the X in $X \to Y$, and the third column is the Y in $X \to Y$.

1.0 [”DAI93865”, ”FRO40251”]

0.999176276771005 [”GRO85051”, ”FRO40251”]

0.9906542056074766 [”GRO38636”, ”FRO40251”]

0.9905660377358491 [”ELE12951”, ”FRO40251”]

0.9867256637168141 [”DAI88079”, ”FRO40251”]

0.983510011778563 [”FRO92469”, ”FRO40251”]

0.972972972972973 [”DAI43868”, ”SNA82528”]

0.9545454545454546 [”DAI23334”, ”DAI62779”]

0.7326649958228906 [”ELE92920”, ”DAI62779”]

0.717948717948718 [”DAI53152”, ”FRO40251”]

Answer to Question 2(e)

Top 10 pairs by confidence (only 5 needed):

The first column is the confidence score, the second column is the X in $X, Y \to Z$, the third column is the Y in $X, Y \to Z$ and the fourth column is the Z in $X, Y \to Z$.

1.0 [”DAI23334”, ”ELE92920”, ”DAI62779”]

1.0 [”DAI31081”, ”GRO85051”, ”FRO40251”]

1.0 [”DAI55911”, ”GRO85051”, ”FRO40251”]

1.0 [”DAI62779”, ”DAI88079”, ”FRO40251”]

1.0 [”DAI75645”, ”GRO85051”, ”FRO40251”]

1.0 [”ELE17451”, ”GRO85051”, ”FRO40251”]

1.0 [”ELE20847”, ”FRO92469”, ”FRO40251”]

1.0 [”ELE20847”, ”GRO85051”, ”FRO40251”]

1.0 [”ELE26917”, ”GRO85051”, ”FRO40251”]

1.0 [”FRO53271”, ”GRO85051”, ”FRO40251”]

Answer to Question 3(a)

We have a total of n rows with m 1s.

$\therefore$ Total number of columns possible with n rows and m 1s $= \binom{n}{m}$

Let's say that we select first k rows out of all these columns for min-hashing. Then the number of columns which have first k rows as all 0 is: $\binom{n-k}{m}$

$\therefore$ The probability of getting a don't know as a min-hash value is: $\frac{\binom{n-k}{m}}{\binom{n}{m}}$

Solving this yields:

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

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

$= \frac{n-k}{n}\frac{n-k-1}{n-1} \ldots \frac{n-k-m+1}{n-m+1} \ldots$ m terms

$= \frac{n-k-m+1}{n}$

$< (\frac{n-k}{n})^m$

Hence proved!

Answer to Question 3(b)

We know that $(1 - \frac{1}{x})^x \approx \frac{1}{e} \ldots$ (1)

We also know that $(1 - \frac{k}{n})^m =$ Probability of don't know

$\ldots$ (2)

We want (2) to be $\frac{1}{e^{10}}$

Now, raising both left and right hand sides to power 10, we have:

$(1 - \frac{1}{x})^{10x} \approx \frac{1}{e^{10}} \ldots$ (3)

Comparing (2) and (3), we have m corresponding to $10x$ and $\frac{n}{k}$ corresponding to x.

$\therefore m = \frac{10n}{k}$

$\therefore k = \frac{10n}{m}$

$\therefore$ We need k to be at least $\frac{10n}{m}$ so that the probability of don't know is at most e^{-10}.

Answer to Question 3(c)

As an example, let one column be $[0, 1, 0, 0]^T$ and the other column be $[0, 1, 1, 1]^T$.

As we can see from these 2 columns, the Jaccard Similarity is simple $\frac{1}{3}$.

If we start at first 2 rows for random cyclic permutation, then both the columns have similar min-hash values. However, if we start at last 2 rows for random cyclic permutation, then both the columns have different min-hash values. Hence, the probability that a random cyclic permutation yields the same min-hash value for both the columns is $\frac{1}{2}$.

Hence, for these 2 columns, the probability that their min-hash values agree is not the same as their Jaccard Similarity.