CSE 547 HW1

Dongyang Wang

Jan 11 2023

Academic Integrity We take academic integrity extremely seriously. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom they interacted with.

Discussion Group (People with whom you discussed ideas used in your answers):

Haochen Hu, Songhao Wu, with whom I chatted very briefly for some details of the assignment.

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

https://github.com/maxpumperla/elephas/issues/183
https://stackoverflow.com/questions/62198911/not-able-to-run-pyspark-in-google-colab
https://stackoverflow.com/questions/21138751/spark-java-lang-outofmemoryerror-java-heap-space/22742982

I acknowledge and accept the Academic Integrity clause.

(Signed) Dongyang Wang

Answer to Question 1

My solution uses the Spark to a limited extent, such that I used it to build the pipeline and processed the data, after which I converted it to a dictionary to allow faster processing. I then generated the results using lists and dictionaries, where I get friends' friends and exclude the ones who are already friends.

The recommendation results are

('924', ['11860', '15416', '2409', '43748', '439', '45881', '6995']),

('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)

Under the independence of A and B, by the statistic property of Independence, the confidence will be $conf(A \rightarrow B) = P(B|A) = P(B)$ Ignoring $P(B)$ might give us a biased result because in the independence case the quantity we desire is solely decided by $P(B)$.

$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{P(B|A)*N}{Support(B)} = \frac{P(B \cap A)}{P(A)P(B)}$. This takes into account the $P(B)$, not missing out information on B (support of B) or A as defined, "A and B occur together".

Similarly, $conv(A \rightarrow B) = \frac{1-S(B)}{1-conf(A \rightarrow B)} = \frac{1-P(B)}{1-\frac{P(B \cap A)}{P(A)}}$ also takes into account the $P(B)$, not missing out information on B (support of B) or A, as defined "probability that A appears without B if they were independent".

Answer to Question 2(b)

Proof. $conf(A \to B)$ is asymmetrical. $conf(A \to B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$ and $conf(B \to A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$. When $P(A) \neq P(B)$, $conf(A \to B) \neq conf(B \to A)$. But when $P(A) = P(B)$, $conf(A \to B) = conf(B \to A)$.

□

Proof. $lift(A \to B)$ is symmetrical. $lift(A \to B) = \frac{conf(A \to B)}{S(B)} = \frac{P(B|A) * N}{Support(B)} = \frac{P(B \cap A)}{P(A)P(B)}$ and $lift(B \to A) = \frac{conf(B \to A)}{S(A)} = \frac{P(A|B) * N}{Support(A)} = \frac{P(B \cap A)}{P(A)P(B)}$.

Therefore, $lift(A \to B) = lift(B \to A)$.

□

Proof. $conv(A \to B)$ is asymmetrical. $conv(A \to B) = \frac{1 - S(B)}{1 - conf(A \to B)} = \frac{1 - P(B)}{1 - \frac{P(B \cap A)}{P(A)}} = \frac{P(A) - P(A)P(B)}{P(A) - P(B \cap A)}$ and $conv(B \to A) == \frac{1 - S(A)}{1 - conf(B \to A)} = \frac{1 - P(A)}{1 - \frac{P(A \cap B)}{P(B)}} = \frac{P(B) - P(A)P(B)}{P(B) - P(B \cap A)}$.

When $P(A) = P(B)$, $conv(A \to B) = conv(B \to A)$. But in general, they are not equal. A counter-example is $Basket1 = \{A, B, C\}, Basket2 = \{A, B\}, Basket3 = \{A\}, Basket4 = \{A, B\}, Basket5 = \{C\}$. $P(A) = 4/5, P(B) = 3/5, P(A \cap B) = 2/5$. So, $conv(A \to B) = \frac{4/5 - 12/25}{2/5} = 4/5$ and $conv(B \to A) = \frac{3/5 - 12/25}{1/5} = 3/5$ which are not equal.

Therefore, when $P(A) \neq P(B)$, $conv(A \to B) \neq conv(B \to A)$.

□

Answer to Question 2(c)

Confidence $conf(A \to B) = \frac{P(A \cap B)}{P(A)} = 1$ for perfect implication since $P(A) = P(B) = P(A \cap B) = 1$ in this case, which is the largest it can get since it's a probability and is between 0 and 1. Therefore, it is desirable.

Conviction $conv(A \to B) = \frac{1 - S(B)}{1 - conf(A \to B)}$ and in this case by statement above we know the denominator is approaching 0 and therefore $conv(A \to B)$ approaches infinity, as large as it can get (while numerator not equal to 0). Therefore, it is desirable.

Lift $lift(A \to B) = \frac{conf(A \to B)}{S(B)} = \frac{P(B|A) * N}{Support(B)} = \frac{P(B \cap A)}{P(A)P(B)} = 1/2$ under perfect implications since since $P(A) = P(B) = P(A \cap B) = 1$. This, however, may not be the largest it can get. For example, when $P(A) = 1, P(B) = 1/2, P(A \cap B) = 1/2$, $lift(A \to B) = 1$ and indeed gets larger. Therefore, it is not desirable.

Answer to Question 2(d)

Table 1: Answer to Question 2(d)

X	Y	Confidence
DAI93865	FRO40251	1.0
GRO85051	FRO40251	0.999176276771005
GRO38636	FRO40251	0.9906542056074766
ELE12951	FRO40251	0.9905660377358491
DAI88079	FRO40251	0.9867256637168141

Answer to Question 2(e)

Table 2: Answer to Question 2(e)

X	**Y**	**Confidence**
('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

Answer to Question 3(a)

Proof. There are $\binom{n}{k}$ options for selecting rows and $\binom{n-m}{k}$ options for selecting rows without any 1's. $P(Don't know) = \frac{\binom{n-m}{k}}{\binom{n}{k}} = \frac{(n-)!k!(n-k)!}{k!(n-m-k)!n!} = \frac{(n-k)(n-k-1)...(n-m-k+1)}{n(n-1)(n-2)...(n-m+1)}$ Since there are m terms and each term is smaller than or equal to $\frac{n-k}{n}$, therefore $P(Don't know) \leq (\frac{n-k}{n})^m$.

□

Answer to Question 3(b)

Proof. From part 1, we want to have $(\frac{n-k}{n})^m \leq e^{-10}$. By approximation, $(\frac{n-k}{n})^m = (1-k/n)^m = (1-\frac{k}{n}^{n/k})^{km/n}$. By definition of e and n is much larger than k and m, $(1-\frac{k}{n}^{n/k})^{km/n} = e^{-km/n}$. Thus, $\frac{km}{n} \leq 10$ such that $k \leq \frac{10n}{m}$ is the quantity we want to ensure this probability is at most e^{-10}.

□

Answer to Question 3(c)

Part a: $S_1 = \{0, 0, 1, 1\}$ and $S_2 = \{0, 1, 0, 1\}$.

Part b: Their Jaccard similarity is 1/3 because among the matching rows with 1's, only (1,1) forms a matching pair.

Part c: Since there is only one case, (1,1), that makes the min-hashing the same, rotating 4 times will result in only 1 in 4 cases of yielding the same min-hash value. So, the probability is 1/4.