CSC4008 - Assignment 2

Yohandi

November 7, 2023

1 Products You Might Buy [80 pts]

Before we delve deeper, $\texttt{Pr}(B) = \frac{\texttt{support}(B)}{N}$.

(a) A drawback of using confidence is that it ignores $\texttt{Pr}(B)$. Why is this a drawback? Explain why lift and interest do not suffer from this drawback.

The confidence measure can be misleading as it does not take into consideration the base probability of the occurrence of B. If B is a very common item (meaning that $\texttt{Pr}(B)$ is high), the confidence measure will always give a high value for any A no matter what A is. This will mislead us into concluding that there are strong associations between A and B.

Lift and interest metrics take into account the base probability of B ($\texttt{Pr}(B)$), which provides a more balanced view of the strength of the association between items A and B. This inclusion helps to prevent the misleading conclusions that can arise when using confidence alone, especially in cases where B is a very common item across transactions.

- **Lift**: By dividing the confidence by $\texttt{Pr}(B)$, lift adjusts for the frequency of B. This normalization allows for a comparison of the observed frequency of A and B occurring together against the frequency expected under independence. If B is a common item, its high probability would make the denominator in the lift calculation larger, which could lower the lift value, hence indicating that the association might not be particularly strong. This ensures that lift is not biased by the high occurrence rate of B.
- **Interest**: This metric compares the confidence of the rule to the expected confidence, assuming that A and B are independent. The expected confidence is simply $\texttt{Pr}(B)$, the probability of seeing B in any transaction. By subtracting this from the actual confidence, the interest value indicates whether the rule is better (positive interest) or worse (negative interest) than what would be expected by chance.

(b) A measure is symmetrical if $\texttt{measure}(A \to B) = \texttt{measure}(B \to A)$. Which of the measures presented here are symmetrical? For each measure, please provide either a proof that the measure is symmetrical, or a counterexample that shows the measure is not symmetrical.

- Confidence is not symmetrical. The confidence of $A \to B$ can be very different from the confidence of $B \to A$. Derivations are as follows:
  - $\texttt{confidence}(A \to B) = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(A)}$
  - $\texttt{confidence}(B \to A) = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(B)}$

  One can simply find any A and B such that $\texttt{Pr}(A) \neq \texttt{Pr}(B)$. A counterexample follows.

  Suppose the probability of a customer buying an apple (A) is 0.8, buying bread (B) is 0.2, and buying both apple and bread ($A \cap B$) is 0.1. Then:

  - $\texttt{confidence}(A \to B) = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(A)} = \frac{0.1}{0.8} = \frac{1}{8}$
  - $\texttt{confidence}(B \to A) = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(B)} = \frac{0.1}{0.2} = \frac{1}{2}$

It is shown that $\texttt{confidence}(A \to B) \neq \texttt{confidence}(B \to A)$.

- Lift is symmetrical. The formula for lift does not change if we swap A and B since it is the ratio of the joint probability of A and B to the product of the probabilities of A and B, which remains the same when A and B are interchanged. Derivations are as follows:
  - $\texttt{lift}(A \to B) = \frac{\texttt{confidence}(A \to B)}{\texttt{Pr}(B)} = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(A)\texttt{Pr}(B)}$
  - $\texttt{lift}(B \to A) = \frac{\texttt{confidence}(B \to A)}{\texttt{Pr}(A)} = \frac{\texttt{Pr}(B \cap A)}{\texttt{Pr}(B)\texttt{Pr}(A)} = \frac{\texttt{Pr}(A \cap B)}{\texttt{Pr}(A)\texttt{Pr}(B)}$

  It is shown that $\texttt{lift}(A \to B) = \texttt{lift}(B \to A)$.
- Interest is not symmetrical. The interest for $A \to B$ can be different from the interest for $B \to A$, since the interest measure depends on the confidence, which we've already established is not symmetrical.

  A counterexample:

  Suppose the probability of a customer buying an apple (A) is 0.8, buying bread (B) is 0.2, and buying both apple and bread ($A \cap B$) is 0.1. Then:
  - $\texttt{interest}(A \to B) = \texttt{confidence}(A \to B) - \texttt{Pr}(B) = \frac{1}{2} - \frac{1}{5} = \frac{3}{10}$
  - $\texttt{interest}(B \to A) = \texttt{confidence}(B \to A) - \texttt{Pr}(A) = \frac{1}{8} - \frac{4}{5} = -\frac{27}{40}$

  Clearly, we have shown that $\texttt{interest}(A \to B) \neq \texttt{interest}(B \to A)$.

(c) Identify pairs of items (X, Y) such that the support of $\{X, Y\}$ is at least 100. For all such pairs, compute the confidence scores of the corresponding association rules: $X \Rightarrow Y$, $Y \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Break ties, if any, by lexicographically increasing order on the left hand side of the rule.

```
Top 5 rules for pairs X => Y are listed below.
DAI93865 => FRO40251
GRO85051 => FRO40251
GRO38636 => FRO40251
ELE12951 => FRO40251
DAI88079 => FRO40251
```

(d) Identify item triples (X, Y, Z) such that the support of X, Y, Z is at least 100. For all such triples, compute the confidence scores of the corresponding association rules: $(X, Y) \Rightarrow Z$, $(X, Z) \Rightarrow Y$, $(Y, Z) \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Order the left-hand-side pair lexicographically and break ties, if any, by lexicographical order of the first then the second item in the pair.

```
Top 5 rules for triples (X, Y) => Z are listed below.
(DAI23334, ELE92920) => DAI62779
(DAI31081, GRO85051) => FRO40251
(DAI55911, GRO85051) => FRO40251
(DAI62779, DAI88079) => FRO40251
(DAI75645, GRO85051) => FRO40251
```

2 Min-hashing [10 pts]

(a) Compute the minhash signature for each column if we use the following three hash functions to simulate the random row permutation. For example, if we use h_1, $h_1(0) = 1$, which means row 0 is mapped to row 1 after the permutation.

1. $h_1(x) = 2x + 1 \bmod 6$

Elements	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Figure 1: Input Matrix

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

2. $h_2(x) = 3x + 2 \texttt{ mod } 6$
3. $h_3(x) = 5x + 2 \texttt{ mod } 6$

(b) Which of these hash functions produce true permutations? Here, a true permutation means that the numbers generated by the hash function for row numbers form a permutation of 0 to 5.

Notice that for $x = [0, 1, 2, 3, 4, 5]$:

- $h_1(x) = [1, 3, 5, 1, 3, 5]$
- $h_2(x) = [2, 5, 2, 5, 2, 5]$
- $h_3(x) = [2, 1, 0, 5, 4, 3]$

We may conclude that only h_3 produces true permutations as the numbers generated by it for row numbers form a permutation of 0 to 5.

3 Estimation via Sampling [10 pts]

(a) Estimate the fraction of students who have taken at least 5 courses.

To ensure unbiased sampling, we must avoid any skew that might be caused by selecting tuples based on attributes that are not uniformly distributed. Since studentID is unique within each university, it is a good candidate for the key in the hash function to guarantee a uniform sample among students.

To construct the sample:

- Hash the studentID to b buckets (where b=20 for a 1/20th sample size).
- Select the tuples if their hash value falls within the first bucket.
- The key attributes in the tuples should include studentID and courseID.
- Count the number of courses per student within the sample.
- To estimate the fraction of students who have taken at least 5 courses, scale up the counts obtained in the sample by the inverse of the sampling ratio (i.e., by 20).

(b) Estimate the fraction of courses where at least half the students got "A"

For this estimation, the courseID can be hashed because we're interested in the course level information. However, since courses are only unique within a university, we have to use both universityID and courseID as a composite key for the hash function to ensure uniform sampling of courses across different universities.

To construct the sample:

- Hash the composite key universityID + courseID into b buckets.
- Select the tuples if their composite key hash value falls within the first bucket.
- From the selected tuples, determine the fraction of students with an "A" grade for each course.
- Estimate the fraction of all courses where at least half the students got an "A" by scaling the counts from the sample by the sampling ratio.
- The key attributes here are universityID, courseID, and grade. After identifying the courses where at least half the students received an "A" within the sample, we would scale that number by the inverse of the sampling fraction to estimate the fraction of all courses that meet this criterion.