Massive Data Analysis

Instructor: Dr.Gholampour

Fall 2024

HW 2

Arman Yazdani - 400102255

[1]You can find practical's section report on *descriptions* of *MDA2024_HW2,ipynb*

1

1.1 *lift*

The probability of A and B occurring with each other, assuming that A and B are independent.

1.2 *conv*

Conviction criterion describes the strength of an association rule. Numerator, possibility of not having item set B and in The denominator is the probability of not having item B under the condition of item set A. So the Numerator shows item set B being rare and the denominator measures item set A and B being unrelated . For example, if item set B is rare but *confidence*($A->B$) is high, *Conviction* increases, which indicates a strong association rule.

1.3 *confidence*

The main weakness of *confidence* is that it does not take into account the probability of B's occurrence.For example, suppose B is a frequent item set that exists in most baskets.In this case, the probability of finding B in the basket containing A is high. That is, in most baskets that have A, B is also present because of its abundance. Hence, $\frac{\Pr(A \cap B)}{\Pr(A)}$ will have a high value. While A and B do not have an association rule.

On the other hand, if B is a rare item set, it will appear in few baskets of A But the reason is the rarity of B, not the absence of an association rule between A and B. This is the weakness of the confidence criterion.

2

2.1 Frequent Items

Items $I_i \in \{1:20\}$ are frequent because they appear in at least 5 baskets, while the rest of the items don't.

2.2 Frequent Item pairs

Item pair can be interpreted to product of Items ($P_{i,j} = I_i \times I_j$).Therefor , Item pairs whos product is in frequent Items($I_i \in \{1:20\}$) are frequent: {{1,{2:10}},{2,{3:10}},{3,{4:6}},{4,5}}

2.3 Sum of all basket sizes

Number	Multiplicants Count	Number	Multiplicants Count
1	100	2	50
3	33	4	25
5	20	6	16
7	14	8	12
9	11	10	10
⋮	⋮	⋮	⋮
99	1	100	1

Total sum = 482

2.4 Samlpe confidence & interest

- {6,9} ⇒ 3:

$$conf(\{6,9\} \Rightarrow 3) = \frac{\Pr(6,9,3)}{\Pr(6,9)} = 1$$
$$intr(\{6,9\} \Rightarrow 3) = conf(\{6,9\} \Rightarrow 3) - \Pr(3) = \tfrac{2}{3}$$

- {2,4,5} ⇒ 3:

$$conf(\{2,4,5\} \Rightarrow 3) = \frac{\Pr(2,4,5,3)}{\Pr(2,4,5)} = \tfrac{1}{3}$$
$$intr(\{2,4,5\} \Rightarrow 3) = conf(\{2,4,5\} \Rightarrow 3) - \Pr(3) = 0$$

3

3.1 Memory for *triangular-matrix*

- We have I Items,therefor our item pairs will be:

$$totalpairs = \binom{I}{2} = \tfrac{I(I-1)}{2}$$

- Each element occupies 4 bytes:

$$totalbytes = 4 \times \tfrac{I(I-1)}{2} = 2I(I-1)$$

3.2 Non-Zero Pairs

There are 2 possible scenarios:

1. All in one basket: in this case all items appear in one basket which makes all possible non-zero pairs:

$$\text{max \# of non-zero pairs} = \binom{I}{2} = \tfrac{I(I-1)}{2}$$

2. One in each basket: this time, we have one basket for each item:

$$\text{max \# of non-zero pairs} = 0$$

3.3 Triples vs Triangular

The space used in the Triangular matrix method was mentioned in section A. This value is constant for storing all possible pairs between 1 item because all pairs are stored even if they do not appear in the data ($Mem_{triangular} = 2I(I-1)bytes$) In the Triples method, only non-zero pairs of pairs that have been observed in the data are stored. Each pair is stored as a triple (count), which includes the index of two items and the number of times that pair has been observed, so the space consumption of this method is as follows compared to the previous method (bytes):

$$\text{Number of pairs} \times 12 = \text{Memorytriples}$$

The number 12 is actually the multiplication of 3 by 4 bytes, which makes a total of 12 bytes In order for this method to occupy less space than the previous method, it should be

$$Mem_{triples} < Mem_{triangular}$$
$$\#ofpairs \times 12 < 2I(I-1)$$
$$\#ofpairs < \frac{I(I-1)}{6}$$

The above result is the condition of occupying less space in this method, if the data is scattered and most of the pairs do not appear. Triples method is better and will consume less space. But if the data is dense and most pairs are observed, the Triangular matrix method will consume less space.

4

4.1 minhash signature

By applying each of h_is to *element* we get(if collision:lowest empty row):

$$h_1^* = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 & 1 \\ 2 & 1 & 0 & 0 & 1 \\ 4 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 5 & 0 & 1 & 1 & 0 \end{bmatrix} \quad h_2^* = \begin{bmatrix} 3 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 4 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 & 1 \end{bmatrix} \quad h_3 = \begin{bmatrix} 2 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 1 & 0 \\ 4 & 1 & 1 & 0 & 1 \\ 3 & 0 & 0 & 1 & 1 \end{bmatrix}$$

finding *minhash* we have:

$$h_1 = [0\ 3\ 0\ 1]\ h_2 = [2\ 1\ 0\ 0]\ h_2 = [0\ 1\ 1\ 0]$$

4.2 Valid Hash function

A permutation hash function is valid if it produces a unique value function for all possible values. In other words, the hash function should be one by one and span so that all possible values in the output interval appear exactly once in the output, according to the stated characteristic

and definition, only h_3 is valid because only in this function whose outputs are unique, the outputs were examined in section A with *.

4.3 *Jaccard*

To calculate the real Jaccard similarity, we must use the original matrix in the question. Also, minhash signature matrix should be used to calculate the estimated Jaccard similarity. so We compare and check each column in pairs:

Actual Jacquard	Estimated Jacquard	Columns
$\frac{3}{3} = 1$	$\frac{1}{5} = 0.2$	S_1, S_2
$\frac{2}{3} = 0.67$	$\frac{1}{6} = 0.17$	S_1, S_3
$\frac{3}{3} = 1$	$\frac{2}{4} = 0.5$	S_1, S_4
$\frac{2}{3} = 0.67$	$\frac{2}{5} = 0.4$	S_2, S_3
$\frac{3}{3} = 1$	$\frac{1}{5} = 0.2$	S_2, S_4
$\frac{2}{3} = 0.67$	$\frac{1}{6} = 0.17$	S_3, S_4

Vivdly, the real and estimated similarity are very different. This is because that only 3 hash functions have been used for estimation, which is a small number. We can do the following to improve the estimation:

- Using a larger number of hash functions allows a more accurate estimate of the Jaccard similarity to be obtained.
- Ensuring that selective hash functions produce unique and valid permutations.
- Reducing the number of elements of sets can also increase the accuracy of estimation.