Assignment 2

# Q1

1. **Confidence** evaluates the likelihood of item B being purchased when item A is bought. However, it overlooks the popularity of item B, which can lead to misleading associations. For instance, if B is a popular item, confidence in (A -> B) may be high, even if no significant connection exists between A and B.

**Lift** and **Interest**, successfully address this constraint. Lift accounts for Pr(B) by calculating the ratio of actual to predicted confidence. Interest directly subtracts the anticipated support for B from the observed confidence, mitigating the influence of B's popularity. By offering a more impartial viewpoint on item associations, these metrics lessen the possibility of drawing erroneous inferences based only on item popularity.

1. **Confidence** is not symmetrical since it is directional.

* *conf(A -> B) = Pr(BlA)* and *conf(B -> A) = Pr(A|B)*. *Pr(AlB)* and *Pr(BlA)* might be different.

Suppose we have two products, A and B,

* + Product A has 100 out of 200 customers who bought A also bought B.
  + Product B has 20 out of 50 customers who bought B also bought A.

Confidence of A -> B *(conf(A -> B)) = Support(A and B) / Support(A)* is 100/200 = 0.5, indicating that 50% of A buyers also bought B.

Confidence of B -> A *(conf(B -> A)) = Support(A and B) / Support(B)* is 20/50 = 0.4, indicating that 40% of B buyers also bought A.

Hence, conf(A -> B) is not equal to conf(B -> A), showing that confidence is not symmetrical. The direction of the association matters in this measure.

**Lift** is symmetrical.

* With proof of its symmetry:
  + *Lift (A -> B) = Pr(A, B) / [Pr(A) \* Pr(B)]*
  + *Lift (B -> A) = Pr(B, A) / [Pr(B) \* Pr(A)]*

Show that *Lift (A -> B) = Lift (B -> A):*

By the commutative property of multiplication, *Pr(A, B)* is the same as *Pr(B, A),* and *Pr(A)* is the same as *Pr(A)*.

Therefore, *Lift (B -> A) = Pr(A, B) / [Pr(B) \* Pr(A)]* Since *Pr(A, B)* and *Pr(B, A)* are the same, and *Pr(A)* and *Pr(A)* are also the same, we have: *Lift (B -> A) = Lift (A -> B)*

This shows that *Lift (A -> B)* is equal to *Lift (B -> A)*, confirming the symmetry of the lift measure. The order of items A and B in the lift measure does not affect the result, making it a symmetrical measure.

**Interest** is not symmetrical.

* *Interest (A -> B) = conf(A -> B) - Pr(B)*
* *Interest (B -> A) = conf(B -> A) - Pr(A)*

Suppose we have two products, A and B:

* + Product A: 60 out of 100 customers who bought A also bought B. So, *conf(A -> B)* = 0.6.
  + Product B: 20 out of 50 customers who bought B also bought A. So, *conf(B -> A)* = 0.4.

Now, let's calculate the Interest measures without making assumptions about *Pr(B)* and *Pr(A):*

* *Interest (A -> B) = conf(A -> B) - Pr(B) = 0.6 - Pr(B)*
* *Interest (B -> A) = conf(B -> A) - Pr(A) = 0.4 - Pr(A)*

In this counterexample, we do not assume specific values for *Pr(B) and Pr(A),* and we observe that *Interest (A -> B)* is not necessarily equal to *Interest (B -> A),* confirming that Interest is not symmetrical. The direction of the association is still important.

1. **Top Pair Rule 1**: ['DAI93865'] => ['FRO40251'], Confidence: 1.0

**Top Pair Rule 2**: ['GRO85051'] => ['FRO40251'], Confidence: 0.9991762767710051

**Top Pair Rule 3**: ['GRO38636'] => ['FRO40251'], Confidence: 0.9906542056074765

**Top Pair Rule 4**: ['ELE12951'] => ['FRO40251'], Confidence: 0.9905660377358491

**Top Pair Rule 5**: ['DAI88079'] => ['FRO40251'], Confidence: 0.9867256637168142

1. **Top Triple Rule 1**: ['DAI23334', 'ELE92920'] => ['DAI62779'], Confidence: 1.0

**Top Triple Rule 2**: ['DAI31081', 'GRO85051'] => ['FRO40251'], Confidence: 1.0

**Top Triple Rule 3**: ['DAI55911', 'GRO85051'] => ['FRO40251'], Confidence: 1.0

**Top Triple Rule 4**: ['DAI62779', 'DAI88079'] => ['FRO40251'], Confidence: 1.0

**Top Triple Rule 5**: ['DAI75645', 'GRO85051'] => ['FRO40251'], Confidence: 1.0

# Q2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Element** | **S1** | **S2** | **S3** | **S4** | **2x+1**  **mod 6** | **3x+2**  **mod 6** | **5x+2**  **mod 6** |
| 0 | 0 | 1 | 0 | 1 | 1 | 2 | 2 |
| 1 | 0 | 1 | 0 | 0 | 3 | 5 | 1 |
| 2 | 1 | 0 | 0 | 1 | 5 | 2 | 0 |
| 3 | 0 | 0 | 1 | 0 | 1 | 5 | 5 |
| 4 | 0 | 0 | 1 | 1 | 3 | 2 | 4 |
| 5 | 1 | 0 | 0 | 0 | 5 | 5 | 3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(0)** | \ | 1 | \ | 1 |
| **h2(0)** | \ | 2 | \ | 2 |
| **h3(0)** | \ | 2 | \ | 2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(1)** | \ | 1 | \ | 1 |
| **h2(1)** | \ | 2 | \ | 2 |
| **h3(1)** | \ | 1 | \ | 2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(2)** | 5 | 1 | \ | 1 |
| **h2(2)** | 2 | 2 | \ | 2 |
| **h3(2)** | 0 | 1 | \ | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(3)** | 5 | 1 | 1 | 1 |
| **h2(3)** | 2 | 2 | 5 | 2 |
| **h3(3)** | 0 | 1 | 5 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(4)** | 5 | 1 | 1 | 1 |
| **h2(4)** | 2 | 2 | 2 | 2 |
| **h3(4)** | 0 | 1 | 4 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **S1** | **S2** | **S3** | **S4** |
| **h1(5)** | 5 | 1 | 1 | 1 |
| **h2(5)** | 2 | 2 | 2 | 2 |
| **h3(5)** | 0 | 1 | 4 | 0 |

The final minhash signature matrix is:

|  |  |  |  |
| --- | --- | --- | --- |
| **S1** | **S2** | **S3** | **S4** |
| 5 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |
| 0 | 1 | 4 | 0 |

* 1. Only function h3(x)=5x+2mod6 is a true permutation, since the numbers generated by the hash function for row numbers form a permutation of 0 to 5.

# Q3

1. To estimate the fraction of students who have taken at least 5 courses in a context where 'studentID' is only unique within a university,

**Key Attribute**: The key attribute used to construct the sample is **'universityID'** in combination with **'studentID'**, since ‘**studentID’** alone is not sufficient for unique identification, the composite key ***[university, studentID])*** is essential to distinguish each record in the dataset.

Explanation:

* Create a stratified sample by independently selecting 1/20th of the tuples for each university.
* Within each university, randomly select a sample of students (e.g., 1/20th of the students) using their **‘studentID’**.
* This sample construction avoids biased estimation.

1. To estimate the fraction of courses where at least half the students received an "A" grade, we must create a composite key using both **'universityID'** and **'courseID'** to uniquely identify each tuple. Since **'courseID'** alone is not sufficient for unique identification, the composite key ***[universityID, courseID]*** is essential to distinguish each record in the dataset.

Explanation:

* Create a stratified sample by independently selecting 1/20th of the tuples for each university.
* Within each university, take a random sample of courses (e.g., 1/20th of the courses) based on their **'courseID'**.
* This sample construction avoids biased estimation.