Association Rule

It is a type of learning technique that checks for dependency of one data item on another data item and maps accordingly, so that it can be more profitable.

Representation

→ If items "coffee" are purchased, then the Possibility of buying "Sugar" is ... which can represented by:

→ $\{\text{Coffee}\} \rightarrow \{\text{Sugar}\}$

ARL (Unsupervised)

- Apriori
- Eclat
- F-P Growth Algorithm.

Working : Support. "People who bought also bought."
Confidence
lift.

Support ⇒ frequency of item (x)

$$\text{Support} = \frac{\text{frequency } (x)}{\text{Total transaction}} \%$$

*** Support count = frequency of x.

Confidence = it indicates how often the items x and y both occurs together in dataset. when the the occurance of x in given.

$$\text{Confidence } (x \rightarrow y) = \frac{\text{freq } (x, y)}{\text{freq } (x)}$$

lift ⇒ it is the strength of any value rule.

$$= \frac{\text{Support}(X, Y)}{\text{Support}(X) \cdot \text{Support}(Y)}$$

or,

$$= \frac{\text{Confidency}(X \to Y)}{\text{Support}(Y)}$$

if, lift = 1, item independent to each other.

if, lift > 1, item dependent to each other.

if lift < 1, one item is substitute for other items.

Application

→ Market Basket analysis

→ Medical Diagnosis.

→ Protein Sequence.

Example

Step 1: Set a minimum support and confidence.

Step 2: Take all the subsets in transactions having higher support than minimum support.

Step 3: Take all the Rules of these subsets having higher confidence than minimum confidence.

Step 4: Sort the rules by decreasing lift.

Transaction id	Item Sets
T_1	A; B
T_2	B, D
T_3	B, C
T_4	A, B, D
T_5	A, C
T_6	B, C
T_7	A, C
T_8	A, B, C, E
T_9	A, B, C

Given, minimum Support = 2

minimum confidence = 50%

Apply Apriori algorithm

⇒ Calculate C_1 ← Candidate set.

L_1 ← frequent item set.

C_1 =

Item set	Support count
A	6
B	7
C	6
D	2
E	1

L_1 =

Item set	Support count
A	6
B	7
C	6
D	2

all the itemset have greater or equal than minimum support

C_2 =

Item set	Support count
{A, B}	4
{A, C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

L_2 =

Item Set	Support count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

C_3 =

Item Set	Support Count
{A, B, C}	2
{A, B, D}	1
{B, C, D}	0
{A, C, D}	0

l_3 =

Item Set	Support Count
{A, B, C}	2

finding Association Rules for subset's :

→ we will calculate the confidence $\frac{Sup(A,B)}{Sup(A)}$

→ we will excludes rules that have less confidence than min threshold (50%)

Rules	Support	Confidence
$\{A, B\} \rightarrow C$	2	$\frac{Sup(\{A, B\}, C)}{Sup(\{A\}, B)} = \frac{2}{4} = 50\%$
$\{B, C\} \rightarrow A$	2	$= \frac{2}{4} = 50\%$
$\{A, C\} \rightarrow B$	2	$= \frac{2}{4} = 50\%$

$C \rightarrow (A, B)$ 2 $\frac{2}{6} = 33.33\%$

$A \rightarrow (B, C)$ 2 $\frac{2}{6} = 33.33\%$

$B \rightarrow (A, C)$ 2 $\frac{2}{7} = 28\%$

So the first 3 Rules $(A, B) \rightarrow C$

$(B, C) \rightarrow A$

$(A, C) \rightarrow B$

Can be consider as the strong association Rules for given problem.

→ Related algorithm questions → checks ML A-Z Course Pdf.

Institute of Information Technology
Noakhali Science and Technology University

Course Code: CSE-4213 *[Class Test-2]* Course Title: Applied Data Science
Duration: 30 Minutes Full Marks: 15

Answer all the following questions

1. Consider the transaction dataset given as follows.

TID	Items
T1	*I1, I2, I5*
T2	*I2, I4*
T3	*I2, I3*
T4	*I1, I2, I4*
T5	*I1, I3*
T6	*I2, I3*
T7	*I1, I2, I3, I5*
T8	*I1, I3*
T9	*I1, I2, I3*

(a) Generate association rules for the items where minimum support count is 2 and minimum confidence is 60%. 3

(b) Calculate the value of *confidence ({I2, I3} => {I1}) and confidence ({I1} => {I2, I3})*. 2

2. Dr. Strange have applied hierarchical clustering and got two clusters, $C1 = \{(1,1), (2,2), (3,1)\}$ and $C2 = \{(5,3), (6,4)\}$. Now he wants to approximate the distance between these two clusters. [Note: Dr. Strange will apply Euclidean distance as it is his favorite distance metric.]

(a) Calculate the distance between $C1$ and $C2$ applying single link distance measure. 1
(b) Calculate the distance between $C1$ and $C2$ applying average link distance measure. 1
(c) Calculate the distance between $C1$ and $C2$ applying complete link distance measure. 1
(d) Calculate the distance between $C1$ and $C2$ applying centroid based distance measure. 1

3. A perceptron can learn the operation *OR* but not Exclusive-OR (*XOR*).

(a) Why do you think a perceptron cannot learn the *XOR* operation whether it learns the *OR* operation? 3

(b) How can you cope with the problem and propose a way so that *XOR* operation can be learnt? 3

Good Luck!

Exam Question Solution

C_1

Itemn	Support count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

L_1

Itemn	Support Count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

C_2

Items	Support count
(I_1, I_2)	4
(I_1, I_3)	4
(I_1, I_4)	1
(I_1, I_5)	2
(I_2, I_3)	4
(I_2, I_4)	2
(I_2, I_5)	2
(I_3, I_4)	0
(I_3, I_5)	1
(I_4, I_5)	0

L_2

(I_1, I_2)	4
(I_1, I_3)	4
(I_1, I_5)	2
(I_2, I_3)	4
(I_2, I_4)	2
(I_2, I_5)	2

$\underline{\underline{C_3}}$

$(I_1, I_2, I_3) \longrightarrow 2$

$(I_1, I_2, I_4) \longrightarrow 1$

$(I_1, I_2, I_5) \longrightarrow 2$

$(I_2, I_3, I_4) \longrightarrow 0$

$(I_2, I_3, I_5) \longrightarrow 1$

$(I_2, I_4, I_5) \longrightarrow 0$

(

$\underline{\underline{L_3}}$

$(I_1, I_2, I_3) \longrightarrow 2$

$(I_1, I_2, I_5) \longrightarrow 2$

$\underline{\underline{C_4}}$

$(I_1, I_2, I_3, I_4) \longrightarrow 0$

$(I_1, I_2, I_3, I_5) \longrightarrow 1$

$(I_1, I_2, I_4, I_5) \longrightarrow 0$

Rules	Support	Confidence
$(I_2, I_3) \rightarrow I_1$	2	$\Rightarrow \dfrac{2}{4}$ $= 50\%$
$(I_1) \rightarrow (I_2, I_3)$	2	$= \dfrac{2}{6}$ $= 33.33\%$