Lec-7 Recommender Systems - 1

- Walmart Case study
- Apriori Algorithm
- Market Basket Analysis
- Association Rules
- Association Metrics

Dataset link:

- - -

Dataset

invoice no.	Item_name	Quantity
1	A	2
1	B	6
1	C	5
2	X	
2	Y	

↓ Transform

Invoice no.	A Item 1	B Item 2	C Item 3	X Item 4	Y Item 5	- - -
1	1	1	1	0	0	
2	0	0	0	1	1	
3						
4						
5						

D = set of all the items

$= \{ I_1, I_2, I_3, \ldots, I_n \}$

$\|D\| = n$

$T_1 = \{ I_1, I_2, I_3 \}$

$T_2 = \{ I_4, I_5 \}$

$\vdots$

$T_m =$

$T = \{ T_1, T_2, T_3, \ldots, T_m \}$

↳ set of all transaction.

$|T| = m$

* Usually $|T| > |D|$

* $T_i \subseteq D$

Eg. $T_1 = \{1, 3, 4\}$

$T_2 = \{1, 2, 4, 5\}$

$T_3 = \{1, 3, 5, 6\}$

$T_4 = = \{1, 3, 4, 5\}$

$T_5 = \{1, 3, 6\}$

Q1) frequency of $\{1, 3\} = 4$

Q2) freq. of $\{1, 4\} = 3$

Q3) freq of $\{1, 3, 4\} = 2$

$T_1 = \{1, 3, 4\}$ → List all the subsets.

Subsets = $\{\}, \{1\}, \{3\}, \{4\}, \{1,3\}, \{1,4\}, \{3,4\}, \{1,3,4\}$

} all of the item combination that can be thought of from the T_1

# of subsets = 8

$||S|| = n$ ⇒ how many subsets of S?

Ans. 2^n

$T_1 = \{1, 3, 4\}$ → every item has Yes/No option to be in the subset

Y/N Y/N Y/N

$2 \times 2 \times 2 = 2^3$

Eg. Y Y N ⇒ {1, 3}

Y N N = {1}

$T_1 = \{1, 3, 4\}$

$T_2 = \{1, 2, 4, 5\}$

$T_3 = \{1, 3, 5, 6\}$

$T_4 = = \{1, 3, 4, 5\}$

$T_5 = \{1, 3, 6\}$

NOTE: To recommend items using a frequency based approach, we need to count the frequency of all item combination.

frequency based approach

$\{1, 2\}$	1
$\{1, 3\}$	4
$\{1, 4\}$	3
$\{1, 5\}$	3
$\{1, 6\}$	2

$\{1, 2, 4\} = 1$

$\{1, 2, 3\} = 0$

$\{1, 3, 4\} = 2$

$\{1, 3, 5\} = 2$

Freq threshold ≥ 3

Q) if we have m transaction and n times ?

How many computations will be needed to implement the frequent based approach in the worst scenario ?

Ans. $m \times 2^n$ → very very large

Apriori Algorithm

$T_1 = \{1, 3, 4\}$

$T_2 = \{1, 2, 4, 5\}$

$T_3 = \{1, 3, 5, 6\}$

$T_4 = = \{1, 3, 4, 5\}$

$T_5 = \{1, 3, 6\}$

frequency based approach

{1}	5
{2}	1
{3}	4
{4}	3
{5}	3
{6}	2

{1,2}	1
{1,3}	4
{1,4}	3
{1,5}	3
{1,6}	2

{1, 2, 4} = 1

{1, 2, 3} = 0

{1, 3, 4} = 2

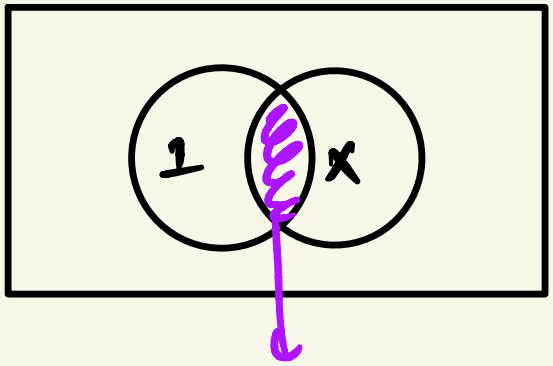
{1, 3, 5} = 2

8) If frequency of {1} is 5, what is the frequency of {1, x}, where x can be element, going to be?

A) > 5

B) < 5 ✓

C) Can't say



{1, x}

{1} is a subset of {1, x}

{1, x} is a superset of {1}

} frequency of subset ≥ freq. of superset

* In apriori, we will set a frequency threshold at the beginning → Eg. freq. threshold = 3

Apriori Algorithm

$T_1 = \{1, 3, 4\}$
$T_2 = \{1, 2, 4, 5\}$
$T_3 = \{1, 3, 5, 6\}$
$T_4 = \{1, 3, 4, 5\}$
$T_5 = \{1, 3, 6\}$

frequency based approach

{1}	5	{4}	3
{2}	1	{5}	3
{3}	4	{6}	2

→ eliminated

⇒ Any combination which contains 2 or 6 will have lesser frequency than this, so they are automatically

{1,3} 4

{1,4} 3

{1,5} 3

{3,4} 2

↳ we don't need to count any combination which has {3,4} in it.

Frequency based approach + Frequency threshold

= Apriori Algorithm

Support

$$\text{Support} = \frac{\text{Frequency of the item combination}}{\text{total \# of transactions}}$$

$T_1 = \{1, 3, 4\}$

$T_2 = \{1, 2, 4, 5\}$

$T_3 = \{1, 3, 5, 6\}$

$T_4 = = \{1, 3, 4, 5\}$

$T_5 = \{1, 3, 6\}$

Support $\{1, 3\} = 4/5 = 0.8$

Support $\{1, 2\} = 1/5 = 0.2$

Support $\{4, 6\} = 0/5 = 0$

Support $\in \{0, 1\}$

Association Rules

1. Bread --> Milk
2. Pizza --> Coke
3. Pizza --> Garlic Bread

Pizza ---> Coke

Antecedent Consequent

if Pizza then Coke

if Antecedent then Consequent

Q) Does the following association rule make sense?

If coke then Pizza → No

⇒ The Association Rules are not symmetric

Association Rules → Association metrics

9) If Bread then Butter

↓ Ante. ↓ Conse.

Which is false → Butter is antecedent

Association Metrics

$X \rightarrow 1000$ $\quad Y \rightarrow 10000$

$X, Y \rightarrow 900$ $\quad X, Y \rightarrow 900$

① Confidence

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

$$= \frac{\#\text{ of transactions containing } X \text{ and } Y \text{ both}}{\#\text{ of transactions containing } X}$$

a) Is $\text{Confidence}(X \dashrightarrow Y) = \text{Confidence}(Y \dashrightarrow X)$

⇒ No. Confidence is asymmetric

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

Eg. if $X \rightarrow Y$ works but not if Y then X

Confidence $(X \rightarrow Y)$ high but Confidence $(Y \rightarrow X)$ low

d) high support, low confidence

low support, high confidence

↳ eg. special cornflakes --> Milk

Lift

* if A and B are independent

$$P(A \cap B) = P(A) * P(B)$$

* if A and B have some positive association, they are not independent.

$$P(A \cap B) \quad \begin{cases} = P(A) * P(B) \\ < P(A) * P(B) \\ > P(A) * P(B) \end{cases}$$

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

e) lift $(X, Y) > 1 \Rightarrow$ +ive correlation or association

$= 1 \Rightarrow$ no correlation

$< 1 \Rightarrow$ -ive correlation

$\text{lift} \in [0, \infty)$

Leverage → Same as lift conceptually

Leverage (X, Y) = Support (X, Y) − Support(X) * Support(Y)

$0 < \text{Leverage}(X, Y) \leq 1$ ⇒ +ive corr.

$\text{Lev}(X, Y) = 0$ ⇒ no corr.

$-1 \leq \text{Lev}(X, Y) < 0$ ⇒ -ive correlation