Agenda

- Business Problem
- Apriori Algorithm
- Association Rule
- Metrics for Association Rule:
  - Support
  - Confidence
  - Lift
  - Leverage
  - Conviction

  } important

Business Problem

Task
- Which products are frequently bought Together

- product placement
- promotion
- Bundles

Reliance Store

Products : Milk, bread, butter - - - - - - - 1000's

$D \in (1, 2, 3, \ldots, n)$

2^3 to 2^{100}

$C_1 \rightarrow T_1 \rightarrow \{1, 3, 8, 10\}$

$C_2 \rightarrow T_2 \rightarrow \{10, 15, 1, 3\}$

$C_3 \rightarrow T_3 \rightarrow \{1, 1000, 10, 3\}$

Pattern ?

- $\{1, 3, 10\}$
- $\{1, 3\}$
- $\{3, 10\}$
- $\{1, 10\}$

1 → recommend 3, 10

1, 3 → 10

Data Set

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01/12/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/10 8:26	3.39	17850.0	United Kingdom

Pre-processing

Invoice No	item1	item2	item3	.....	itemN
101	1	0	1	.....	1
102	0	0	1	.....	1

Sparse Matrix

Task : Find Frequently occuring items

D : Set of all items

T : Set of all Transacti

$T_i \subseteq D$

Apriori Algorithm

Key	Value
{1,2}	10
{1,3,10}	20
{i, y, k ...}	Count

itemsets — Frequency

D ⇒ {1, 2, 3}

itemsets {1}, {2}, {3}, {1,2}
{1,3}, {2,3}, {1,2,3}

D ⇒ {1, 2, 3 ... n} 7

itemsets ? $2^n - 1 - n$

100 ⇒ $2^{100} - 100 - 1$

Brute-force {3}

```
i_sets = get_all_sets(D)
for item_set in i_sets
    for t in transactions:
        if item_set in t:
            Q[itemset] += 1
```

T.C. ∅ 2^n x no-transaction

2^{4000} x 15,000

Very Very Slow

Thresholding

$T = 100$

Ex:

$\{2\} \rightarrow 99$

$\{2,3\} \leq 99$

minimum support

item-set < threshold

↓

remove all supersets of this item-set

#ex2:

$\{4,3\} < T$

$\{2,4,3\}$X $\{2,4,3,10,11\}$X

Bruteforce → $O(2^n \times m)$

↓

reduce this drastically by removing all sets which doesn't satisfy minimum support

* FP Growth (frequent itemset)

↓

TRIES

Market Basket Analysis

Key	Value
{1, 23	10
{1, 3, 103	20
{i, y, k ...}	Count

$(5-6)$ (8) 2^{44} 2^{19}

Applications of Market Basket Analysis

Bio informatics

two chemicals → C_1, C_2

frequently occuring protien across chemicals

Gene sequences → m_1, m_2

ATTC and AGTC

frequently Observed

Medicine → m_1, m_2, m_3 (relationship)

Webpage mining → w_1, w_2, w_3

$\{w_i, w_j, w_k\} \Rightarrow$ frequent

Drawback → Dataset → Millions of Products

Market Basket Analysis with huge number products becomes very costly and time-inefficient

Association Rule

- Quantify the relationship discovered in market Basket Analysis output

$X \Rightarrow \{1, 2, 3\}$

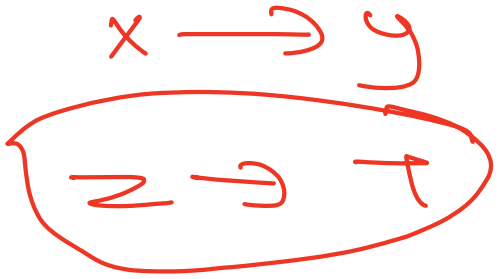
$Y \Rightarrow \{4, 6\}$

$Z \rightarrow X \cup Y$

$Z \rightarrow \{1, 2, 3, 4, 6\}$ → frequent set

$X \longrightarrow Y$

People who buy X have high likelihood of buying Y



q_1

q_2

$q_2 > q_1$

Dominos Pizza

Pizza ⟶ Coke ⇒ rule 1

Pizza ⟶ GarlicBread ⇒ rule 2

rule 1 > rule 2

Coke ⟶ Pizza = Pizza ⟶ Coke

No

X

X ⟶ Y

(Antecedent) ⟶ Consequent

(if) (then)

Bread ⟶ Butter 63%

Ante Consq

Association Metrics

- Support
- Confidence — Conviction
- Lift — Leverage

Support

$$\text{Support}(X) \Rightarrow \frac{\#\text{Transactions}(X)}{\#\text{Total Transactions}}$$

$$\#P(X \text{ in } T)$$

milk = 100

Total = 1000

$$\text{Support}(\text{milk}) = \frac{100}{1000} = 0.1\%$$

Confidence

$$conf(X \rightarrow Y) \Rightarrow \frac{\#(X \text{ and } Y)}{\#(X)}$$

1) $\Rightarrow P(Y/X)$

$P \rightarrow C$ ⊖ $\frac{P \text{ and } C}{P}$

$C \rightarrow P$ ⊖ $\frac{C \text{ and } P}{C}$

$$conf(P \rightarrow C) \; != \; conf(C \rightarrow P)$$

Interpretation

of all the times X occurs
how many times Y also occurs

---

① $\{i+1, i+2\}$ ⇒ High frequency
↓
High Support

$i+1 \rightarrow i+2$ ⇒ low confidence

(2) {it1, it2} → Low frequency
Low support

it1 → it2 → P(it1 / it2) → High Confidence

Lift

Yogurt → milk (High Confidence)

Cornflakes → milk (High Confidence)

Toothpaste → milk ?

Not Sure

* Milk itself is a very frequent item

Transact → 100

Milk → 80

Toothpaste → 14

* Let's say: 10 of them contain both ToP and milk

Confidence (TP → milk) = $\frac{10}{14}$ = 0.7

Lift $(X \rightarrow Y)$ =

$$\frac{\text{Support}(X \cap Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

=

$$\frac{P(X \text{ and } Y)}{P(X) \times P(Y)} = 1$$

Interpretation

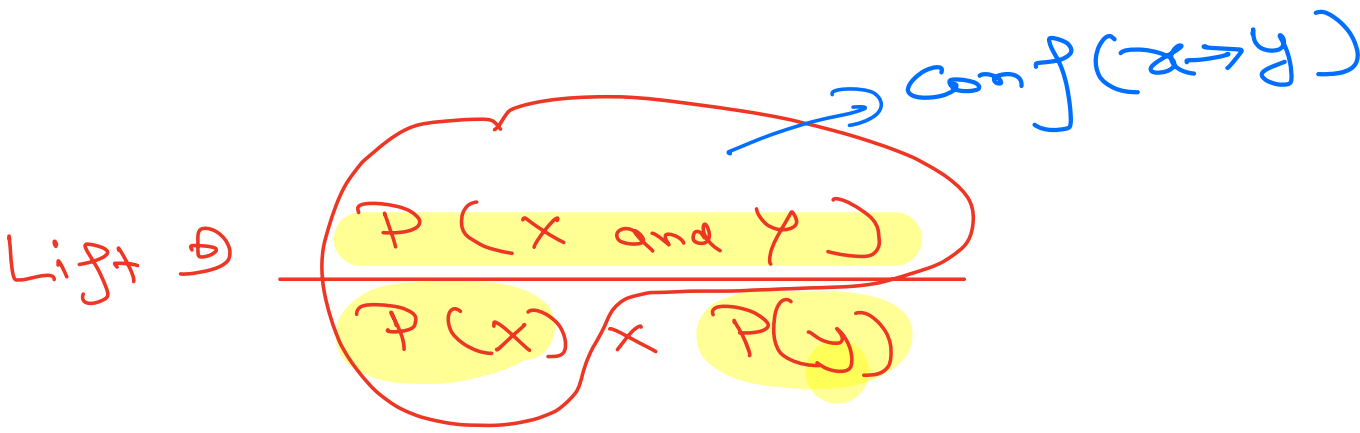
Increase in probability of X and Y together compared to when they are independent.

Lift $(X \rightarrow Y) = 1$; X and Y are independent

Lift $(X \rightarrow Y) > 1$; likely to be bought together

Lift $(X \rightarrow Y) < 1$; Negatively relat

Lift (TP → milk) = $\frac{14/100}{\frac{80}{100} \times \frac{10}{100}}$ = 0.8



$$Lift(X \to Y) = \frac{Conf(X \to Y)}{Support(Y)}$$

$$= \frac{0.7}{0.8} = < 1$$

Leverage → ① Same as Lift ② Easier to interpret

Range: Lift → (0, ∞)

Leverage → (−1, 1)

$$Lev(X \to Y) = Support(X \cap Y) - Support(X) * Support(Y)$$

- Implement of the Metric in Python
- Conviction

Conviction

Ratio of Expected frequency that X occurs without Y if X and Y were independent / Observed frequency of incorrect predictions

$$Conv(X \rightarrow Y) = \frac{1 - s(Y)}{1 - c(X \rightarrow Y)}$$

Note: A High value means Consequent depends Strongly on the antecedant

Q: Can you calculate Lift of Milk and Bread from given data:

- Total Transactions = 600,000
- Transaction {Bread} = 4500
- Transaction {Milk} = 60000
- Transactions {Milk, Bread} = 6000

In Brief

Lift (M → B)

$$\frac{\text{Conf}(M \rightarrow B)}{\text{Support}(B)}$$

Lift (B → M)

$$\frac{\text{Conf}(B \rightarrow M)}{\text{Support}(M)}$$

Cut ⇒ 100

only 10 bags

Milk + Bread > 10

{23 → all customers who bought

↘ ↘