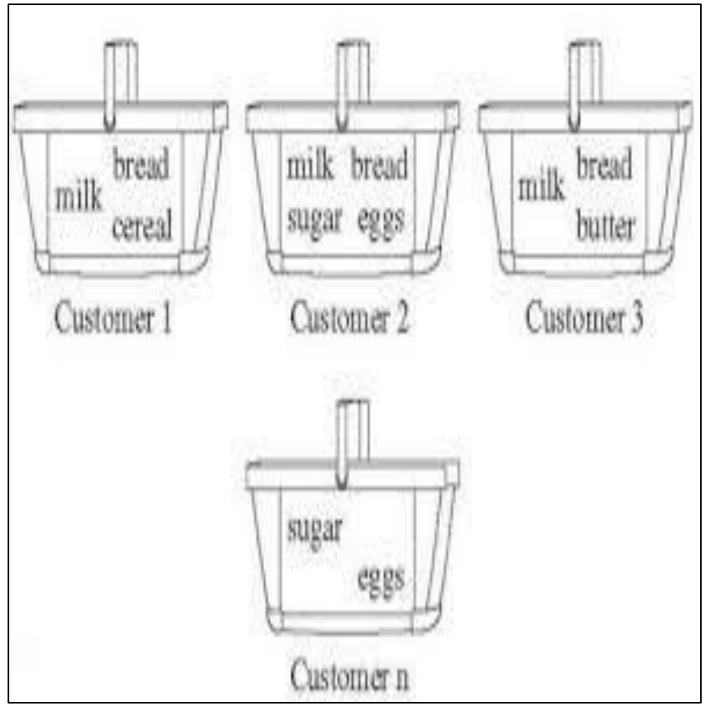
Association Rule Learning

(Introduction, Applications, Basic Concepts, Naïve Algorithm)

CSED, TIET

Association Rule Learning- Introduction

- Association rule learning often known as 'market basket' analysis is very effective technique to find the association of sale of item X with item Y.
- In other words, market basket analysis consists of examining the items in the baskets of shoppers checking out at a market to see what types of items 'go together' (as shown in figure).
- For instance, from the figure it is clear that 'milk' and 'bread' often go together.



Defining Association Rule Mining

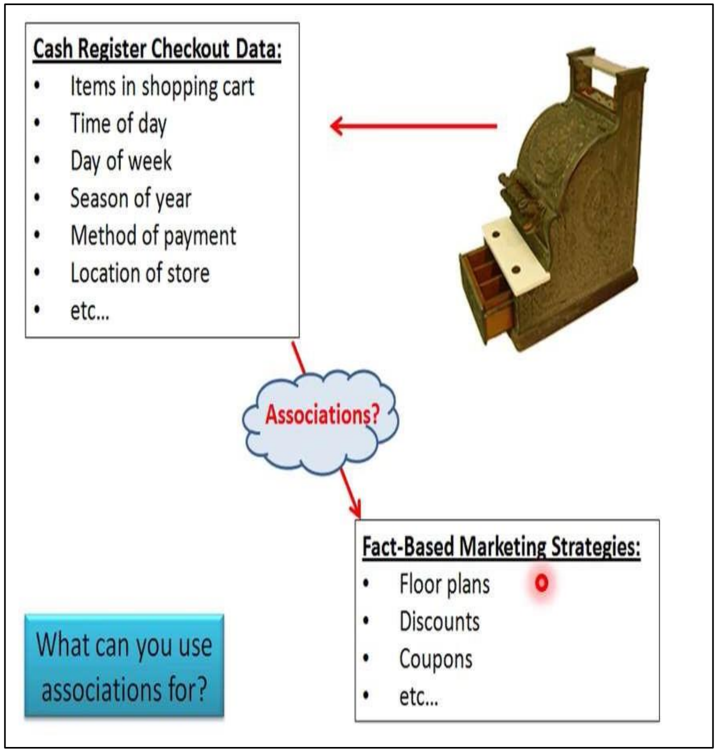
- Association rule mining can be defined as identification of frequent patterns, correlations, associations, or casual structures among sets of objects or items in transactional databases, relational databases, and other information repositories.

- For example, “if a customer buys a dozen eggs, he is 90% likely to also purchase milk”

- These patterns, correlations, associations, or casual structures among sets of objects or items are represented as association rules.

Applications of Association Rule Mining

- Often, while conducting market basket analysis, the only source of data available is the bills of customers.
- The bill tells items in the shopping cart, time of the day, day of the week, season of the year, method of payment, location of store, etc.
- Discovering associations among these attributes can lead to fact-based marketing strategies like:
  - Floor Plans
  - Special Discounts
  - Coupons offerings
  - Product Clustering
  - Catalogue Design
  - Combo packs
  - Customer Segmentation
  - Adaptive Learning
  - Intrusion Detection
  - Web usage mining
  - Bioinformatics



Representation of Items for Association Mining

- Let us assume, that there are n items that a shop stocks.
- For example, there are 6 items in stock, these are Bread, Beer, Eggs, Milk, Cola, Daipers.
- The item list is represented by I and its items are represented as $\{i_1, i_2, i_3, \ldots\ldots, i_n\}$
- The number of transactions are represented by unique identifier TID and each transaction consists of a subsets of items (m<=n) purchased by one customer.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Representation of Items for Association Mining (Contd….)

Transaction ID	Items
T1	I1, I2, I4
T2	I4, I5
T3	I1, I3
T4	I2, I4, I5
T5	I4, I5
T6	I2, I3, I5
T7	I1, I3, I4

Simple Storage

TID	I1	I2	I3	I4	I5
T1	1	1	0	1	0
T2	0	0	0	1	1
T3	1	0	1	0	0
T4	0	1	0	1	1
T5	0	0	0	1	1
T6	0	1	1	0	1
T7	1	0	1	1	0

Horizontal Storage

Item	TID						
	T1	T2	T3	T4	T5	T6	T7
I1	1	0	1	0	0	0	1
I2	1	0	0	1	0	1	0
I3	0	0	1	0	0	1	1
I4	1	1	0	1	1	0	1
I5	0	1	0	1	1	1	0

Vertical Storage

Association Rules

- Association rules are generally represented as ***if/then*** statements.
- An association rule consists of two parts, i.e., an antecedent (if) and a consequent (then).
- An antecedent is an object or item in the data while a consequent is an object or item found in the combination with the antecedent.
- Association rules are often written as $X \rightarrow Y$ meaning that whenever X appears Y also tends to appear.
- X and Y may be single item/object or set of items.
- For example, the rule found in the sales of data of a supermarket could specify that if a customer buys onions and potatoes together, he or she will also like to buy burgers. This rule will be represented as onions, potatoes $\rightarrow$ burgers

Evaluating Association Rule Strength

- Following metrics are used to judge and accuracy of the association rule:

1. Support
2. Confidence
3. Lift

Support

- If there are N transactions in a database, then support of an item X is the number of times X appears in the transactions is to the total number of transactions. i.e., the probability of X.

$$Support\ (X) = \frac{number\ of\ times\ X\ appears\ in\ transactions}{total\ number\ of\ transactions} = \frac{n(X)}{N}$$

- Similarly, support of X and Y is the number of times X and Y both appears in the same transaction is to the total number of transactions. i.e., P(X ∩Y)

$$Support\ (X,Y) = \frac{number\ of\ times\ X\ and\ Y\ both\ appears\ in\ same\ transactions}{total\ number\ of\ transactions} = \frac{n(X \cap Y)}{N}$$

Support -Example

$$Support\ (Bread) = \frac{\mathrm{n(bread)}}{\mathrm{N}} = \frac{4}{5}$$

$$Support\ (Milk) = \frac{\mathrm{n(milk)}}{\mathrm{N}} = \frac{4}{5}$$

$$Support\ (eggs) = \frac{\mathrm{n(eggs)}}{\mathrm{N}} = \frac{1}{5}$$

$$Support\ (Beer, Diaper) = \frac{\mathrm{n(beer \cap diaper)}}{\mathrm{N}} = \frac{3}{5}$$

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Confidence

- Let’s consider that there is a rule X → Y, with a support of 80%, it means that in 80% of times the transactions items X and Y appear together. So it is a rule of interest for the decision/policy makers.
- Let’s say there is another rule A → B, with a support 50%, i.e., the items A and B are not frequent in the transactions and it seems that the rule is not important.
- But it is seen in the transaction data that whenever A appears there is 90% chances that B will appear and hence it is of great interest.
- So, along with support (that measures how frequent items occur in transactions) there is another important metric *Confidence*.

Confidence

- Confidence of a rule X → Y, is the conditional probability of Y when X has been bought.

$$Confidence = P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{n(X \cap Y)}{n(X)} = \frac{Support\ of\ X\ and\ Y}{Support\ of\ X}$$

- Therefore, confidence will be low if X appears much more frequently than X and Y appearing together.
- A high level of confidence shows that the rule is often enough to justify a decision based on it.

Confidence - Example

$$Confidence\ (Bread \rightarrow Milk)$$

$$= \frac{\text{n(Bread,Milk)}}{\text{n(Bread)}} = \frac{3}{4} = 75\%$$

$$Confidence\ (Beer \rightarrow Daipers)$$

$$= \frac{\text{n(Beer,Diaper)}}{\text{n(Beer)}} = \frac{3}{3} = 1 = 100\%$$

$$Confidence\ (Milk, Bread \rightarrow Cola)$$

$$= \frac{\text{n(Milk,Bread,Cola)}}{\text{n(Milk,Bread)}} = \frac{1}{3} = 33.33\%$$

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Alternate representation of rules

- Sometimes, the representation of the rules also includes support and confidence of the rule

$$\text{i.e., } X \rightarrow Y \text{ [Support, Confidence]}$$

- For instance, Beer → Diapers [60%, 100%],

which means that if beer is bought then diapers are bought in 100% of the cases and in 60% of the total rows in the transaction.

Lift

- As already discussed, support of a rule X → Y is number of times X and Y appears is total number of transactions.
- Confidence of a rule X → Y is number of times X and Y appears is to number of times X appear.
- So, both of the metrics consider frequency of X and Y, and confidence consider frequency of X.
- In order to evaluate the strength of the rule, it is also important to consider the frequency of Y.
- For instance, the rule candle → coke [50%, 90%] is not too valid even if it has high confidence. Because coke is a product which is bought with every item. So, whenever candle is bought 90% of times coke is also bought.
- Lift is a metric that considers, joint frequency of X and Y, frequency of X and frequency of Y.

Lift

- Lift of a rule X →Y is defined as ratio of the conditional probability of Y given X is to the probability of Y.
- In other words, lift of a rule X →Y is defined as ratio of confidence of the rule to the probability of Y.

$$Lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{P(Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} = \frac{n(X \cap Y) \times N}{n(X) \times n(Y)}$$

- Therefore, the rule candle → coke [50%, 90%] will have low lift because the frequency of coke in transactions will be quite high.
- The rule candle → matchstick [50%, 90%] will have high lift because the frequency of matchstick is small and hence it is very important rule as it has high confidence and lift.

Lift-Example

$$Lift(Bread, Milk \rightarrow Diapers)$$

$$= \frac{n(Bread, Milk, Diapers) \times N}{n(Bread, Milk) \times n(Diapers)}$$

$$= \frac{2 \times 5}{3 \times 4} = \frac{10}{12} = 83.33\%$$

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Association Rule Mining- Naïve Algorithm

- Naïve Algorithm for Association Rule Mining is a brute force approach to find association rules that satisfy the threshold value of support and confidence.
- It is called **naïve or brute-force technique** because it works by **generating every possible pairs** of items in the transaction database.
- **Following steps are followed for Association Rule Mining:**

1. **Find frequency of each item** and all **possible item pairs**.
2. Identify **item pairs that qualify threshold** value of support.
3. Generate and **identify rules that qualify** the **threshold** value of **confidence**.

Naïve Algorithm- Example

Find association rules with minimum support of 50% and confidence of 75%

Transaction ID	*Items*
100	Bread, Cornflakes
101	Bread, Cornflakes, Jam
102	Bread, Milk
103	Cornflakes, Jam, Milk

Naïve Algorithm Example- Solution

Step 1: Find frequency of each possible item/item pairs.

Itemsets	*Frequency*
Bread	**3**
Cornflakes	**3**
Jam	**2**
Milk	**2**
(Bread, Cornflakes)	**2**
(Bread, Jam)	1
(Bread, Milk)	1
(Cornflakes, Jam)	**2**
(Cornflakes, Milk)	1
(Jam, Milk)	1
(Bread, Cornflakes, Jam)	1
(Bread, Cornflakes, Milk)	0
(Bread, Jam, Milk)	0
(Cornflakes, Jam, Milk)	1
(Bread, Cornflakes, Jam, Milk)	0

Naïve Algorithm Example- Solution (Contd....)

- **Step 2: Identify the item pairs that qualify the threshold value of support.**
- In this case, the threshold value for support is 50%. i.e., select all the item pairs which occur atleast 50% of the total transactions.
- Since, total transactions are 4. So, all item pairs with frequency greater than equal to 2 are accepted (ignoring single items) (highlighted as yellow in figure).

Itemsets	Frequency
Bread	3
Cornflakes	3
Jam	2
Milk	2
(Bread, Cornflakes)	2
(Cornflakes, Jam)	2

Naïve Algorithm Example- Solution (Contd….)

Step3: Generate and identify rules that qualify threshold value of confidence

- Following two item-pairs qualify minimum support threshold:
  - (Bread, Cornflakes)
  - (Cornflakes, Jam)
- So, following four association rules are possible from these item pairs:
  - Bread $\longrightarrow$ Cornflakes
  - Cornflakes $\longrightarrow$ Bread
  - Cornflakes $\longrightarrow$ Jam
  - Jam $\longrightarrow$ Cornflakes

Naïve Algorithm Example- Solution (Contd….)

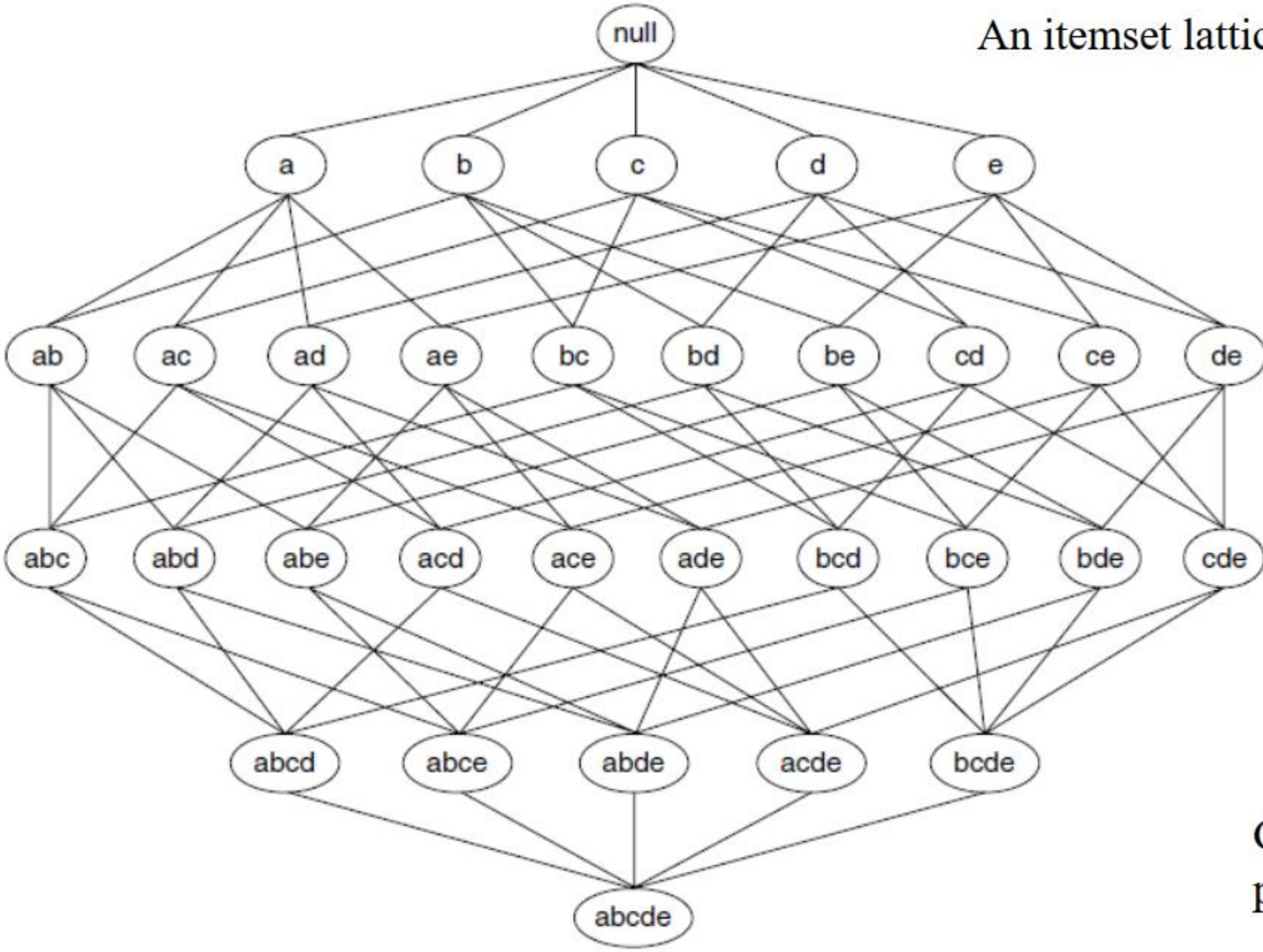
$$Confidence(\text{Bread} \rightarrow \text{Cornflakes}) = \frac{n(\text{Bread} \cap \text{Cornflakes})}{n(Bread)} = \frac{2}{3} = 66.67\%$$

$$Confidence(\text{Cornflakes} \rightarrow \text{Bread}) = \frac{n(\text{Bread} \cap \text{Cornflakes})}{n(Cornflakes)} = \frac{2}{3} = 66.67\%$$

$$Confidence(\text{Cornflakes} \rightarrow \text{Jam}) = \frac{n(\text{Jam} \cap \text{Cornflakes})}{n(Cornflakes)} = \frac{2}{3} = 66.67\%$$

$$Confidence(\text{Jam} \rightarrow \text{Cornflakes}) = \frac{n(\text{Jam} \cap \text{Cornflakes})}{n(Jam)} = \frac{2}{2} = 100\%$$

- Therefore, the only rule Jam → Cornflakes qualifies the minimum confidence threshold.



An itemset lattice

Given d items, there are 2^d possible candidate itemsets

Limitations of Naïve Algorithm

- In Naïve algorithm, we have to find frequency of each possible item pairs by re-iterating the transaction database again and again.
- For n items in the transaction database, Naïve algorithm needs to find frequency of $2^n - n - 1$ (ignoring singleton and null subsets).
- An improved version of Naïve algorithm is to compute only nonzero frequency pairs.
- This can be done by generating item pairs for items that appear in the same transaction (as shown in the figure).

Transaction ID	Items	Combinations
100	Bread, Cornflakes	(Bread, Cornflakes)
200	Bread, Cornflakes, Jam	(Bread, Cornflakes), (Bread, Jam), (Cornflakes, Jam), (Bread, Cornflakes, Jam)
300	Bread, Milk	(Bread, Milk)
400	Cornflakes, Jam, Milk	(Cornflakes, Jam), (Cornflakes, Milk), (Jam, Milk), (Cornflakes, Jam, Milk)