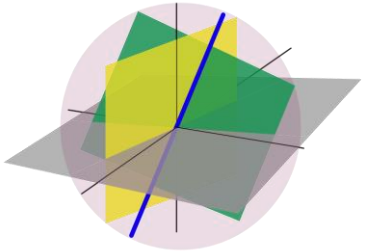
Information Technology Institute



DATA MINING

Machine Learning

Information Technology Institute

Who am I?

Khaled Shaker

BI Consultant – Etisalat &

ITI Teaching Assistant,

MSc Medical Physics.

Information Technology Institute

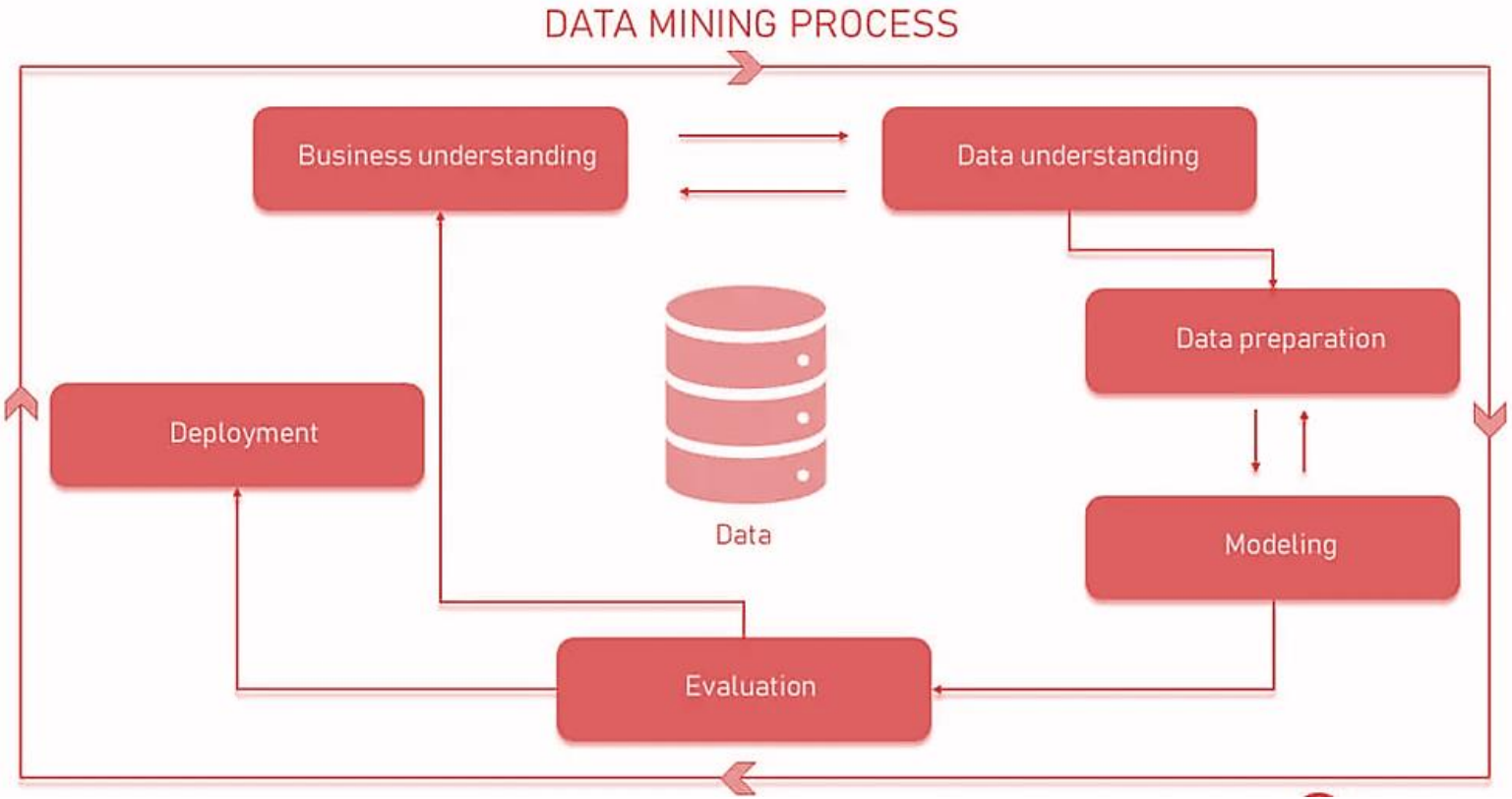
Information Technology Institute

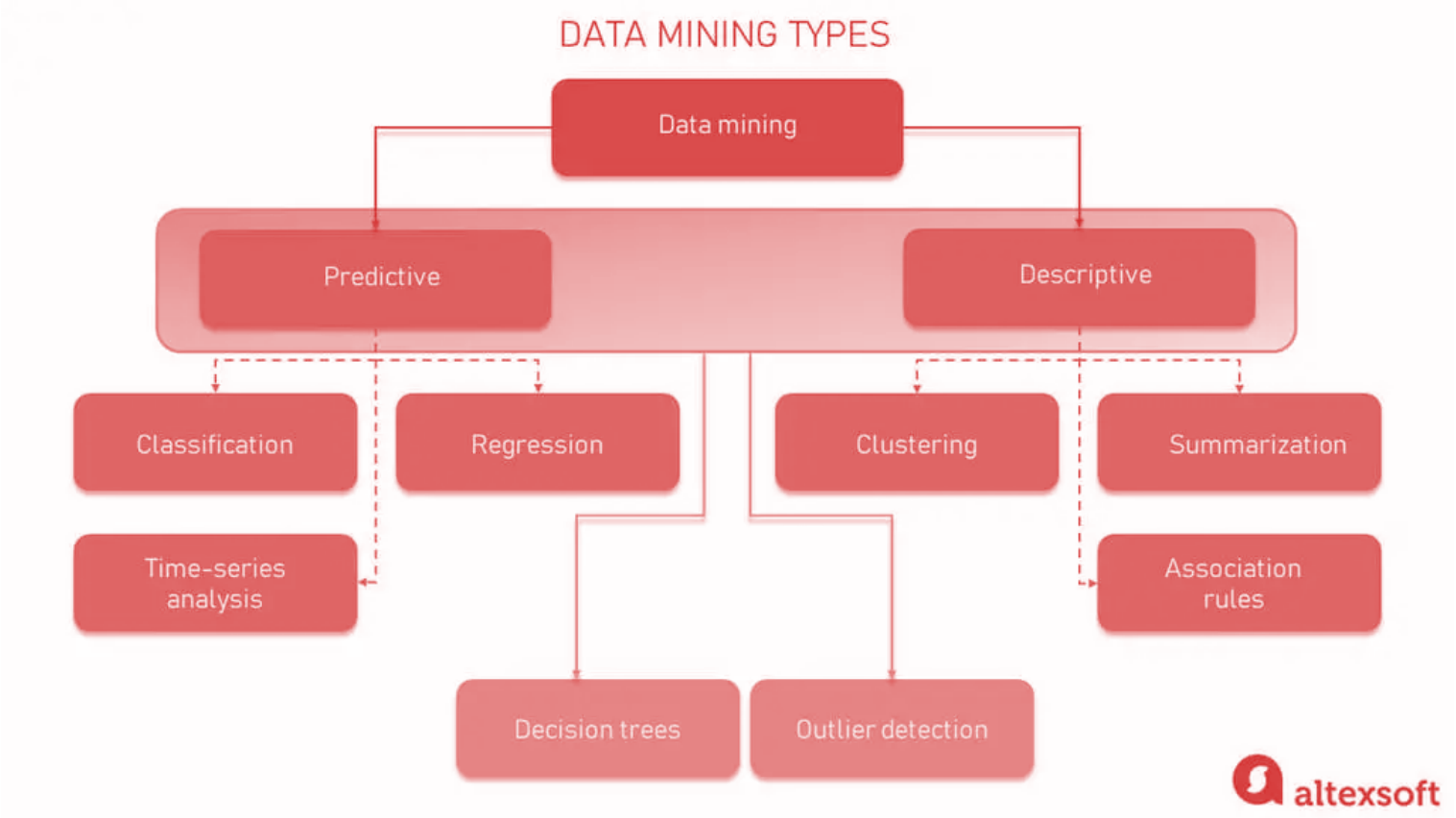
Introduce your self

- Name
- Faculty
- ML prior Experience
- Expectations

What is the data mining?

- Data mining is the process of identifying valid, novel, useful, and understandable patterns in data.
- Also known as KDD (Knowledge Discovery in Databases).
- “We’re drowning in information, but starving for knowledge.” (John Naisbett)


DATA MINING PROCESS
Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data


DATA MINING TYPES
Data mining
Predictive
Descriptive
Classification
Regression
Clustering
Summarization
Time-series analysis
Association rules
Decision trees
Outlier detection
altexsoft

Association Rule

What is Association rule?

Association rule mining is a technique to identify underlying relations between different items.

For instance, if item A and B are bought together more frequently then several steps can be taken to increase the profit. For example:

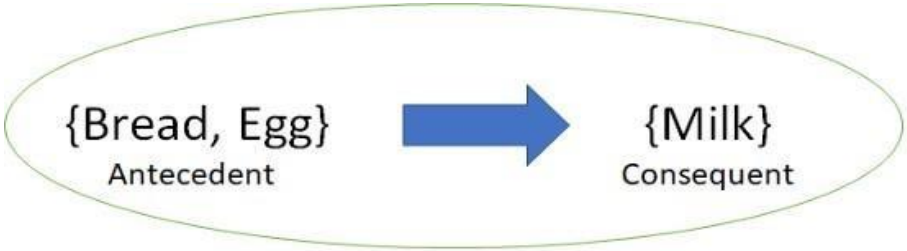
- **A and B can be placed together so that when a customer buys one of the product he doesn't have to go far away to buy the other product.**
- **People who buy one of the products can be targeted through an advertisement campaign to buy the other.**
- **Collective discounts can be offered on these products if the customer buys both of them. Both A and B can be packaged together.**

Customer 1

Customer 2

Customer 3

Customer n



Itemset = {Bread, Egg, Milk}

The main applications of the association rule mining:

- **Basket data analysis** is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.

- **Cross marketing** : is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.

- **Catalog design** : the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another .So these items are often complements or very related.

Apriori Algorithm for Association Rule Mining

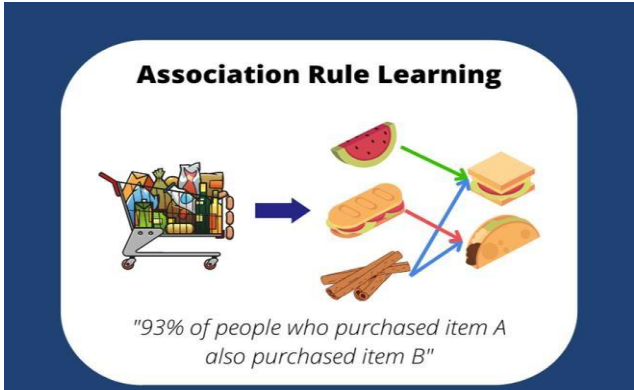
Mining for associations among items in a large database of sales transaction is an important database mining function

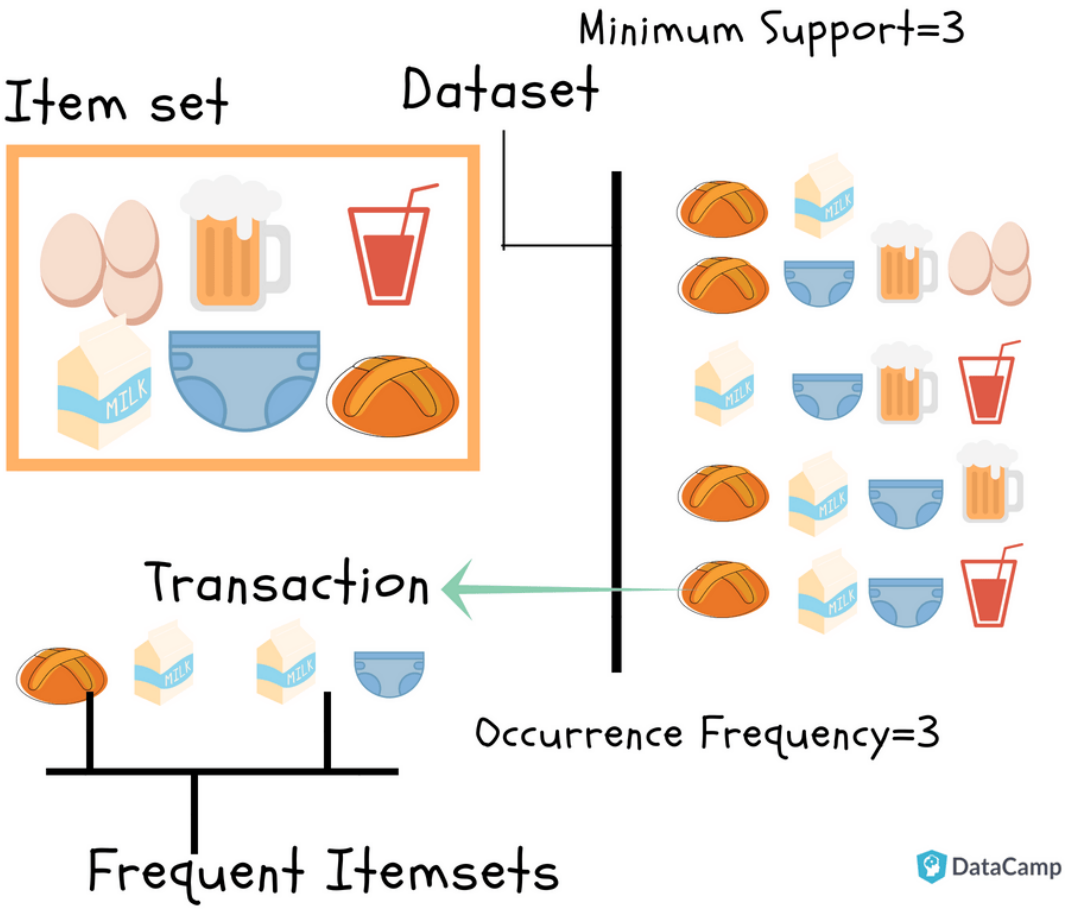
For example, the information that a customer who purchases a keyboard also tends to buy a mouse at the same time is represented in association rule

Keyboard ⇒Mouse
[support = 60%, confidence = 70%]

There are three major components of Apriori algorithm:

- Support
- Confidence
- Lift




Minimum Support=3
Item set
Dataset
MILK
Transaction
Occurrence Frequency=3
Frequent Itemsets
DataCamp

Information Technology Institute

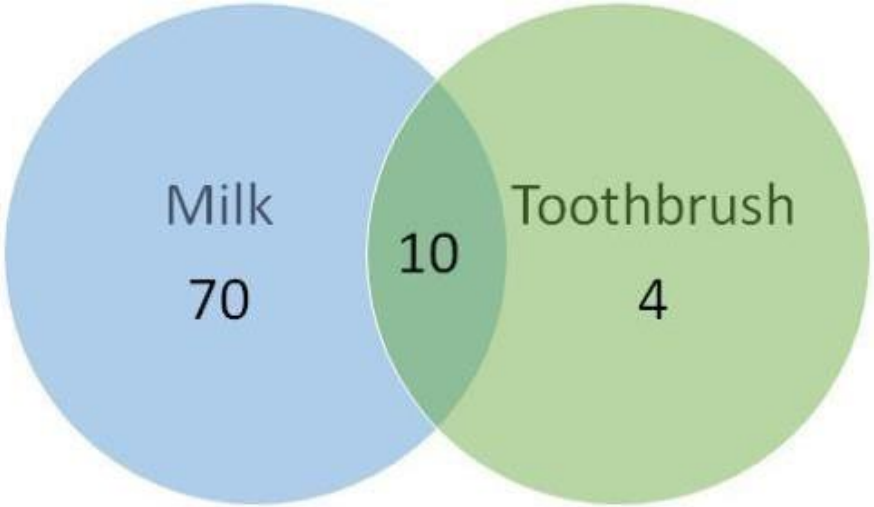
Support: **This measure gives an idea of how frequent an itemset is in all the transactions**

$$Support(X \Rightarrow Y) = \frac{\#\ times\ X\ and\ Y\ occur\ in\ the\ same\ basket}{total\ number\ of\ baskets}$$

Confidence:

- How likely Y is purchased when X is purchased.
- The confidence value indicates how reliable this rule is.

$$Confidence(X \Rightarrow Y) = \frac{\#\ times\ X\ and\ Y\ occur\ in\ the\ same\ basket}{\#times\ X\ occurs\ in\ a\ basket}$$



Consider the numbers from figure on the left. Confidence for {Toothbrush} → {Milk} will be

10/(10+4) = 0.7

Lift: is a measure of importance of a rule.

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Expected\ Confidence}$$

$$Expected\ Confidence = Support(Y)$$

- A lift value greater than 1 indicates that the rule X and the rule Y appear more often together than expected, this means that the occurrence of the rule X has a positive effect on the occurrence of the rule Y.

- A lift smaller than 1 indicates that the rule X and the rule Y appear less often together than expected, this means that the occurrence of the rule X has a negative effect on the occurrence of the rule Y.

- A lift value near 1 indicates that the rule X and the rule Y appear almost as often together as expected, this means that the occurrence of the rule X has almost no effect on the occurrence of the rule Y.

Information Technology Institute

$$Rule: X \Rightarrow Y$$

$$Supprt = \frac{Frequency(X,Y)}{N}$$

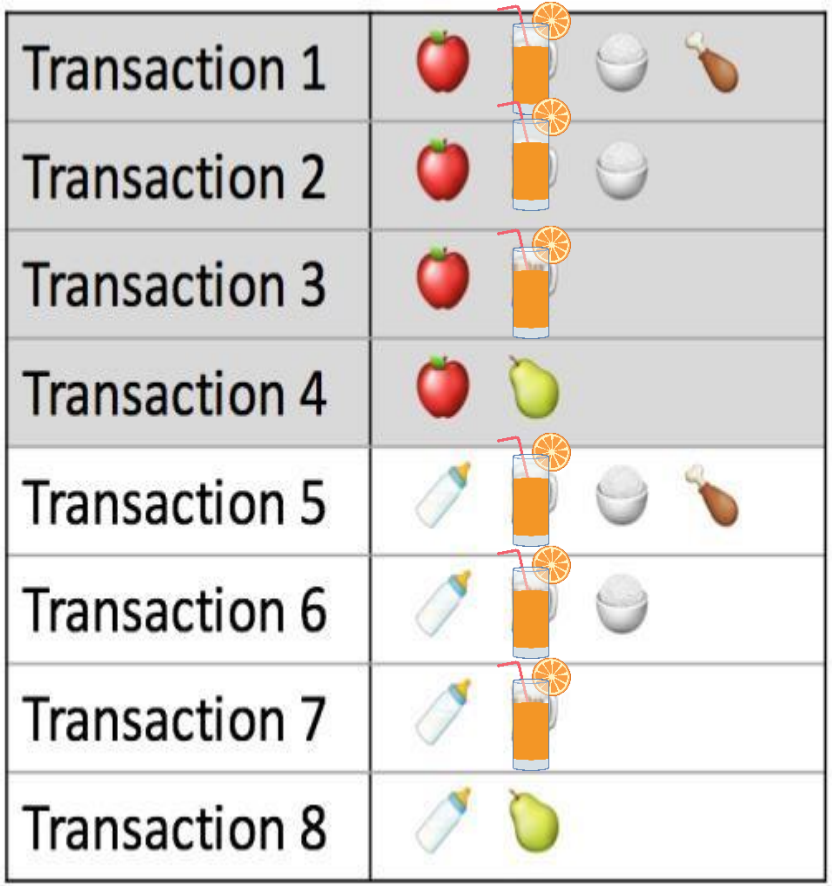
$$Confidence = \frac{Frequency(X,Y)}{Frequency(X)}$$

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

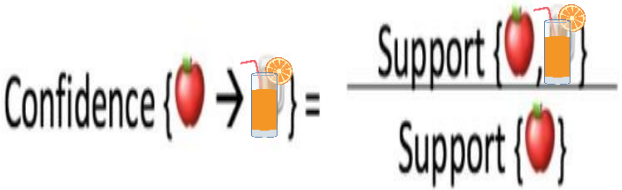
Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Support. This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In Table 1 below, the support of {apple} is 4 out of 8, or 50%. Itemsets can also contain multiple items. For instance, the support of {apple, juice, rice} is 2 out of 8, or 25%.

$$\text{Support } \{🍎\} = \frac{4}{8}$$



Measure 2: Confidence. This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears. In Table 1, the confidence of {apple -> Juice} is 3 out of 4, or 75%.



Transaction 1	
Transaction 2	
Transaction 3	
Transaction 4	
Transaction 5	
Transaction 6	
Transaction 7	
Transaction 8	

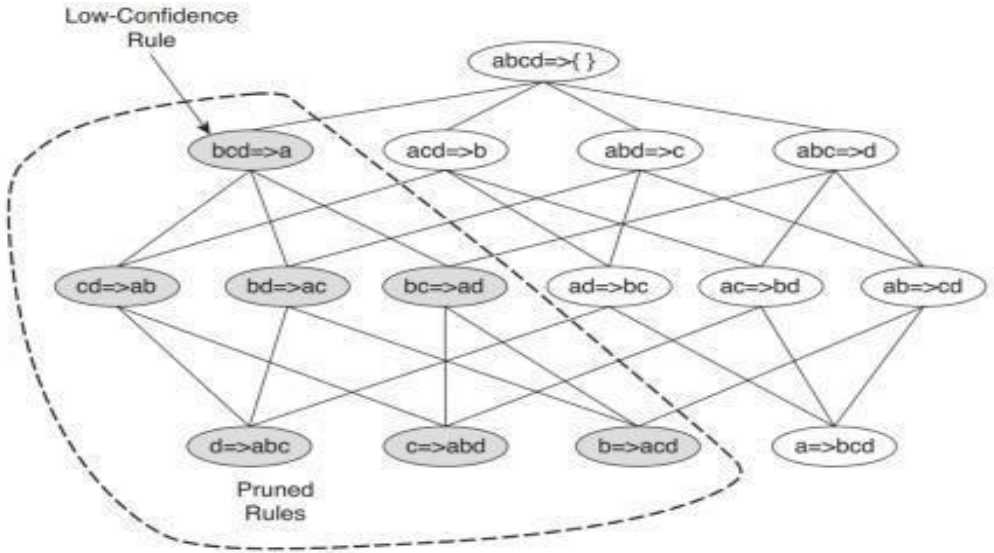
Measure 3: Lift. This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In Table 1, the lift of {apple -> Juice} is 1,which implies no association between items. A lift value greater than 1 means that item Y is *likely* to be bought if item X is bought, while a value less than 1 means that item Y is *unlikely* to be bought if item X is bought.

An Illustration

Transaction 1	🍎 🧃 ⚪ 🍗
Transaction 2	🍎 🧃 ⚪
Transaction 3	🍎 🧃
Transaction 4	🍎 🍐
Transaction 5	🍼 🧃 ⚪ 🍗
Transaction 6	🍼 🧃 ⚪
Transaction 7	🍼 🧃
Transaction 8	🍼 🍐

Steps to solve association rules:

1. **Generating itemsets from a list of items**
2. **Generating all possible rules from the frequent itemsets**



Frequent item set

- Suppose min_sup is the minimum support threshold
- An itemset satisfies minimum support if the occurrence frequency of the itemset is greater or equal to min_sup
- If an itemset satisfies minimum support, then it is a frequent itemset

Strong Rules:

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong

Association Rules:

Mining one level Association (Apriori)

Example:

Assume the following Database transaction:

Transaction	Items
T1	Milk, Bread, Cookies, Juice
T2	Milk, Juice
T3	Milk, Egg
T4	Bread, Cookies, Coffee

With minimum support = 0.5 (2)

Association Rules:

Mining one level Association (Apriori)

Solution:

Step1: Create 1st Level Item set

Item	Support
Milk	3
Bread	2
Cookies	2
Juice	2
Egg	1
Coffee	1

Rejected as they are Below the minimum support

Association Rules:

Mining one level Association (Apriori)

Solution:

Step2: Create 2nd Level Item set

Items	Support
Milk, Bread	1
Milk, Cookies	1
Milk, Juice	2
Bread, Cookies	2
Bread, Juice	1
Cookies, Juice	1

Rejected as they are Below the minimum support

Rejected as they are Below the minimum support

Association Rules:

Mining one level Association (Apriori)

Solution:

Step3: Create 3rd Level Item set

Items	Support
Milk, Juice, Bread	1
Milk, Juice, Cookies	1
Milk, Bread, Cookies	1
Juice, Bread, Cookies	1

Rejected as they are Below the minimum support

There is no association at the 3rd level item set

Association Rules:

Mining one level Association (Apriori)

Solution:

We stop the combination of itemset in one of two cases:

- All the last level items are neglected as they are less than the min support
- Reach Level Item set contains all element

Last Step: Association Rules

Milk=>Juice [support = 0.5, confidence = 0.67]

Juice=>Milk [support = 0.5, confidence = 1]

Bread=>Cookies [support = 0.5, confidence = 1]

Cookies=>Bread [support = 0.5, confidence = 1]

TID	List of Items
1	Beer,Diaper,Baby Powder,Bread,Umbrella
2	Diaper,Baby Powder
3	Beer,Diaper,Milk
4	Diaper,Beer,Detergent
5	Beer,Milk,Coca-Cola

Min_sup 40% (2/5)

C1 → L1

Item	Support
Beer	"4/5"
Diaper	"4/5"
Baby Powder	"2/5"
Bread	"1/5"
Umbrella	"1/5"
Milk	"2/5"
Detergent	"1/5"
Coca-Cola	"1/5"

Item	Support
Beer	"4/5"
Diaper	"4/5"
Baby Powder	"2/5"
Milk	"2/5"

C2 → L2

Item	Support
Beer, Diaper	"3/5"
Beer, Baby Powder	"1/5"
Beer, Milk	"2/5"
Diaper,Baby Powder	"2/5"
Diaper,Milk	"1/5"
Baby Powder,Milk	"0"

Item	Support
Beer, Diaper	"3/5"
Beer, Milk	"2/5"
Diaper,Baby Powder	"2/5"

- C3 → empty

Item	Support
Beer, Diaper,Baby Powder	"1/5"
Beer, Diaper,Milk	"1/5"
Beer, Milk,Baby Powder	"0"
Diaper,Baby Powder,Milk	"0"

So we're going to back to C2

Item	Support(A,B)	Suport A	Confidence
Beer, Diaper	60%	80%	75%
Beer, Milk	40%	80%	50%
Diaper,Baby Powder	40%	80%	50%
Diaper,Beer	60%	80%	75%
Milk,Beer	40%	40%	100%
Baby Powder, Diaper	40%	40%	100%

min_sup=40% min_conf=7

Results

- Juice → Diaper
- support 60%, confidence 70%

- Diaper → Juice
- support 60%, confidence 70%

- Milk → Juice
- support 40%, confidence 100%

- Baby_Powder → Diaper support 40%, confidence 70%

Interpretation

- Some results are belivable, like Baby_Powder → Diaper.
- Some rules need aditional analysis, like Milk → Juice.
- Some rules are unbelivable, like Diaper -> Juice
- This example could contain unreal results because of the small data.

Thank you

khaledgama4@gmail.com

https://www.linkedin.com/in/khaledshakerrr/