Association Rules -> Market Basket Analysis

DSLA COURSE

ROHIT PADEBETTU

Market Basket Analysis

Applications

Market Basket Analysis
Understand customer shopping habits

Default Risk Analysis
Understand which customers are more likely to default

Customer Churn Analysis
Understand which customers are likely to switch

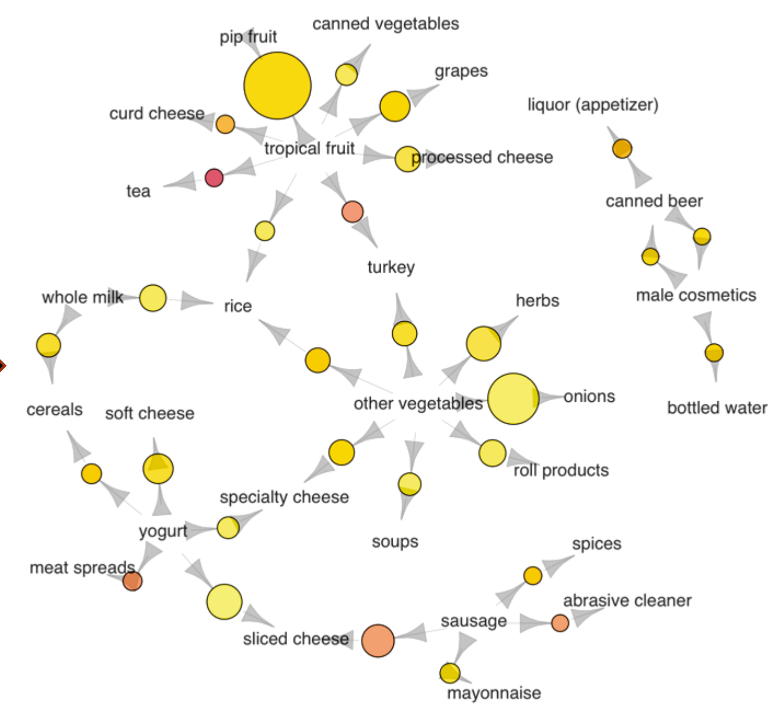
Medical Diagnosis
Helps doctors diagnose illness or even find a treatment

Crime Investigation
Helps investigators understand patterns and associations in crimes

Hurricane Predictions
Helps forecasters identify and predict the intensity of storms and hurricanes

Transactions -> Associations

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐



Support

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Measure of how popular an item is

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

Confidence

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Measures how likely the RHS item is purchased given LHS is purchased

Confidence {🍎 → 🍺} = Support {🍎, 🍺} / Support {🍎}

This measure can tend to inflate and show spurious associations when both LHS and RHS are independently popular

Lift

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Measures how likely the RHS item is purchased given LHS is purchased adjusting for independent popularity of RHS

$$\text{Lift }\{🍎 \rightarrow 🍺\} = \frac{\text{Support }\{🍎, 🍺\}}{\text{Support }\{🍎\} \times \text{Support }\{🍺\}}$$

Association Rules

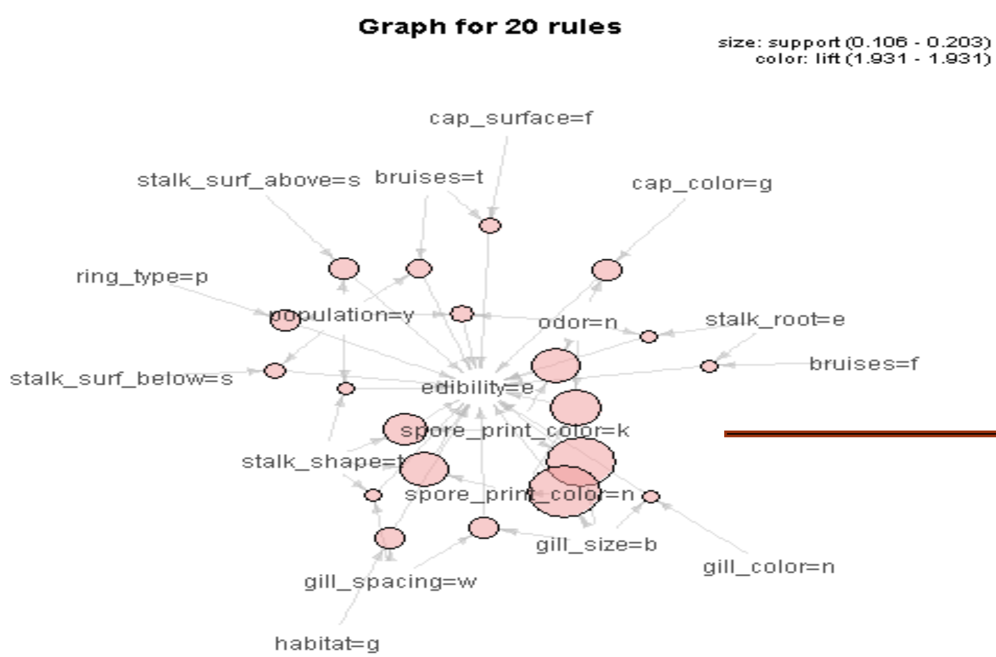
$$Rule: \; X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

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

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

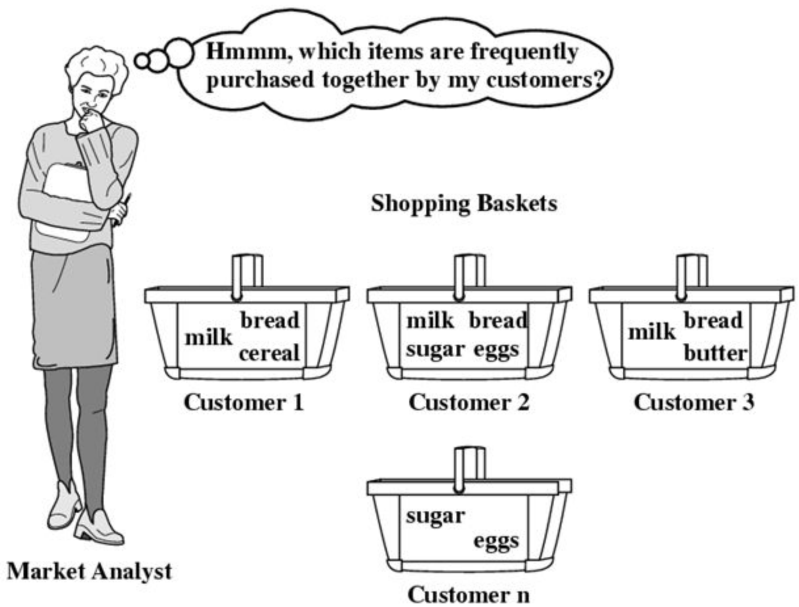
Association Rules - Visualization



Can you guess the dataset ?

Mushroom Dataset !

Interesting Rules



Association Rules are interesting when they satisfy both a minimum support and a minimum confidence

Useful Rules

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Transaction	Support
Canned Beer	10%
Soda	20%
Berries	3%
Male Cosmetics	0.5%

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

Association Rules- Caution

Correlation doesn't imply Causation!

The rules below only suggest a strong co-occurrence relationship between items

Causation requires knowledge about Cause and Effect attributes and typically needs information about how relationships evolve over time

Men who purchase diapers also tend to buy beer at the same time !

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

Association Rules - Computation

Brute Force Approach

- List each item in the basket
- List all possible rules from such items
- Count support and confidence of all such rules
- Prune the rules failing minimum thresholds

N item basket -> 2^N-1 Rules

10 item basket 1023 rules !

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Apriori Algorithm

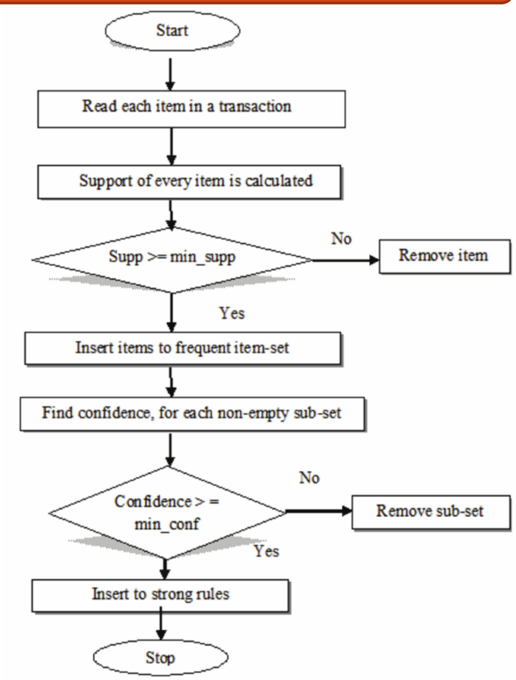
Mathematical Formulation

$$
\begin{array}{l}
\text{Apriori}(T, \epsilon) \\
\quad L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\} \\
\quad k \leftarrow 2 \\
\quad\quad \textbf{while } L_{k-1} \neq \ \ emptyset \\
\quad\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\} \\
\quad\quad\quad \textbf{for } \text{transactions } t \in T \\
\quad\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
\quad\quad\quad\quad \textbf{for } \text{candidates } c \in C_t \\
\quad\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
\quad\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ \ count[c] \geq \epsilon\} \\
\quad\quad\quad k \leftarrow k + 1 \\
\quad\quad \textbf{return } \bigcup_k L_k
\end{array}
$$

Apriori Algorithm

Pseudo Code & Flow Chart

procedure APRIORI_FREQUENTITEMSETS(min_sup, S)
  $L_1 \leftarrow itemsets$
  for $k = 2; L_{k-1} \neq \emptyset; k++$ **do**
    $C_k = aprioriGen(L_{k-1})$ ▷ Create the candidates
    for each $c \in C_k$ **do**
      $c.count \leftarrow 0$
    end for
    for each $I \in S$ **do**
      $C_r \leftarrow subset(C_k, I)$ ▷ Identify candidates that belong to I
      for each $c \in C_r$ **do**
        $c.count++$ ▷ Counting the support values
      end for
    end for
    if $c.count \geq min\_sup$ **then**
      $L_k = L_k \cup c$
    end if
  end for
  return L_k
end procedure



Apriori Algorithm - Principle

RULE1 : *If an "Itemset" is frequent, then all of its subsets must also be frequent*

If {A,B} is frequent, then both {A} & {B} are frequent

RULE2 : *If an "Itemset" is infrequent, then all of its supersets must also be infrequent*

If {A} is infrequent, then {A,B}, {A,C} & {A,B,C} are infrequent

Anti-Monotonicity

These principles are useful to prune candidates.

Apriori Algorithm - Principle

Step 0. Start with itemsets containing just a single item, such as {apple} and {pear}

Step 1. Determine the support for itemsets. Keep the itemsets that meet your minimum support threshold, and remove itemsets that do not

Step 2. Using the itemsets you have kept from Step 1, generate all the possible itemset configurations.

Step 3. Repeat Steps 1 & 2 until there are no more new itemsets

Apriori Algorithm - Example

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Apriori Algorithm - Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Association Rules

Demo

Association Rules

Have good rest of weekend!