CREATIVE INSTITUTE DATA SCIENCE

Association Rule

Association Rule in machine learning is a rule-based method used to discover interesting relationships (i.e., "if-then" patterns) among variables in large datasets.
It is part of unsupervised learning, primarily used in market basket analysis, recommendation systems, and behaviour pattern discovery.

Market Basket Analysis (MBA) is a data mining technique used to discover relationships between products purchased together, helping retailers understand customer behaviour and increase sales.

It analyzes large datasets, like purchase history, to identify frequent product groupings and associations.

Market basket analysis is a strategic data mining technique retailers use to enhance sales by understanding customer purchasing patterns.

This method involves examining substantial datasets, such as historical purchase records, to unveil inherent product groupings and identify items that customers tend to buy together.

For example, if customers are buying milk, how likely are they to also buy bread (and which kind of bread) on the same trip to the supermarket?

This information may lead to an increase in sales by helping retailers do selective marketing based on predictions, cross-selling, and planning their shelf space for optimal product placement.

- **Purpose:** MBA aims to identify patterns in customer buying habits, specifically what products are frequently purchased together.

- **Methodology:** It involves analyzing transaction data to find relationships between different items, such as "customers who buy X also buy Y".

- **Applications:**
  - **Product Placement:** MBA can help determine where to place products on shelves or in online catalogs to encourage combined purchases.
  - **Cross-selling and Upselling:** Identifying products that are commonly bought together can lead to targeted cross-selling and upselling opportunities.
  - **Marketing Campaigns:** Understanding customer purchasing patterns can be used to develop more effective marketing campaigns.
  - **Personalized Recommendations:** MBA can be used to provide personalized product recommendations based on a customer's past purchases.
  - **Inventory Management:** MBA can help optimize inventory by identifying which products are most popular and likely to be sold together.

- **Example:** If MBA analysis reveals that customers who buy diapers often also buy baby wipes, retailers might place these products close together or offer a bundle discount.

Key metrics used in MBA:

- **Support** is measures of how frequently an itemset appears in the dataset.
- **Confidence** is measures of how often B is bought when A is bought.
- **Lift** shows how much more likely A and B are bought together compared to if they were bought independently.

$$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$$

Example

TID	**Items**
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread, Butter
T4	Milk, Bread
T5	Bread, Butter
T6	Milk, Butter

Support of (Bread)

$$\text{Support}(\{Bread\}) = \frac{\text{Number of transactions with Bread}}{6}$$

`Bread` appears in T1, T2, T3, T4, T5 → 5 times.

$$\text{Support}(\{Bread\}) = \frac{5}{6} \approx 0.833$$

Support of (Milk, Butter)

Transactions with both Milk and Butter: T1, T3, T6 → 3 times.

$$\text{Support}(\{Milk, Butter\}) = \frac{3}{6} = 0.5$$

Support of (Milk, Bread, Butter)

Transactions with all three: T1 and T3 → 2 times.

$$\text{Support}(\{Milk, Bread, Butter\}) = \frac{2}{6} = 0.333$$

Confidence

For the rule:

$$X \Rightarrow Y$$

The **confidence** is:

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} = P(Y|X)$$

Confidence of {Milk} → {Butter}

Support(Milk) = Transactions with Milk = T1, T3, T4, T6 = 4

Support(Milk ∩ Butter) = T1, T3, T6 = 3

$$\text{Confidence}(Milk \Rightarrow Butter) = \frac{3}{4} = 0.75$$

Confidence of {Bread} → {Butter}

- Bread appears in: T1, T2, T3, T4, T5 → 5 transactions
- Bread and Butter together in: T1, T2, T3, T5 → 4 transactions

$$\text{Confidence}(Bread \Rightarrow Butter) = \frac{4}{5} = 0.8$$

Confidence = 0.75 for Milk → Butter means:
75% of the times when Milk was bought, Butter was also bought.

Higher confidence = Stronger implication.

Lift

Lift measures how much more likely **item Y** is purchased when **item X** is purchased **compared to Y being purchased independently**.

It helps evaluate whether the rule **X** → **Y** is actually useful or just a coincidence.

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} = \frac{P(X \cap Y)}{P(X) \cdot P(Y)}$$

Interpretation of Lift

Lift Value	Meaning
= 1	X and Y are **independent** (no real association)
> 1	X and Y are **positively correlated** (buying X → Y)
< 1	X and Y are **negatively correlated** (buying X → not Y)

Example

Lift of Milk → Butter

- Support(Milk) = 4/6 = 0.667
- Support(Butter) = 5/6 = 0.833
- Support(Milk ∩ Butter) = 3/6 = 0.5
- Confidence(Milk → Butter) = 0.5 / 0.667 ≈ **0.75**

$$\text{Lift}(Milk \Rightarrow Butter) = \frac{0.75}{0.833} \approx 0.9$$

Or directly using support values:

$$\text{Lift}(Milk \Rightarrow Butter) = \frac{P(Milk \cap Butter)}{P(Milk) \cdot P(Butter)} = \frac{0.5}{0.667 \times 0.833} \approx 0.9$$

Apriori algorithm

The Apriori algorithm is a classic algorithm in data mining used for frequent itemset mining and association rule learning over transactional datasets.

It is based on the Apriori principle:

"If an itemset is frequent, all of its subsets must also be frequent."

This helps reduce the search space and improve efficiency.

Rakesh Agrawal and Ramakrishnan Srikant introduced the method originally in 1994.

The Apriori algorithm, like the preceding example, detects the most frequent itemsets or elements in a transaction database and establishes association rules between the items.

The method employs a “bottom-up” strategy, in which frequent subsets are expanded one item at a time (candidate generation), and groups of candidates are checked against the data.

When no more successful rules can be obtained from the data, the algorithm stops.

How Apriori algorithm work

1. **Association rule:** For example, X Y is a depiction of discovering Y on a basket that contains X.
2. **Itemset:** For example, X,Y is a representation of the list of all objects that comprise the association rule.
3. **Support:** Transactions containing the itemset as a percentage of total transactions
4. **Confidence:** Given X, what is the likelihood of Y occurring?
5. **Lift:** Confidence ratio to baseline likelihood of occurrence of Y

TID	**Items**
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread, Butter
T4	Milk, Bread
T5	Bread, Butter
T6	Milk, Butter

Apriori **step-by-step** using **Support threshold = 0.5 (50%)**

Step 1: Find 1-itemsets (L1)

Count how many times each item appears:

Item	**Support Count**	**Support**	**Frequent? (≥ 0.5)**
Milk	4	4/6 = 0.667	✓ Yes
Bread	5	5/6 = 0.833	✓ Yes
Butter	5	5/6 = 0.833	✓ Yes

L1 = {Milk}, {Bread}, {Butter}

Step 2: Generate 2-itemsets (C2) from L1

Make combinations of frequent items:

- {Milk, Bread}
- {Milk, Butter}
- {Bread, Butter}

Count their support:

Itemset	Support Count	Support	Frequent?
{Milk, Bread}	3	3/6 = 0.5	✓ Yes
{Milk, Butter}	3	3/6 = 0.5	✓ Yes
{Bread, Butter}	4	4/6 = 0.667	✓ Yes

L2 = All 3 pairs

Generate 3-itemsets (C3) from L2

Only combine if all subsets are frequent.

- Candidate: {Milk, Bread, Butter}
    - Subsets: {Milk, Bread}, {Milk, Butter}, {Bread, Butter} → All are frequent

Count support

Itemset	Support Count	Support	Frequent?
{Milk, Bread, Butter}	2	2/6 = 0.333	✗ No

L3 = ∅ (empty) → Stop here.

Final Frequent Itemsets:

- **L1**: {Milk}, {Bread}, {Butter}
- **L2**: {Milk, Bread}, {Milk, Butter}, {Bread, Butter}

Association Rules from Frequent Itemsets:

frequent itemset: {Milk, Butter}

Rule: Milk → Butter

- Support(Milk ∩ Butter) = 3/6 = 0.5
- Support(Milk) = 4/6 = 0.667
- Confidence = 0.5 / 0.667 = **0.75**

Now calculate Lift:

- Support(Butter) = 5/6 = 0.833
- Lift = 0.75 / 0.833 = **~0.9**

Summary Table

Rule	**Support**	**Confidence**	**Lift**
Milk → Butter	0.5	0.75	0.9
Bread → Butter	0.667	0.8	0.96
Milk → Bread	0.5	0.75	0.9

Python Implimentation

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
import matplotlib.pyplot as plt
```

```
dataset = [
    ['Milk', 'Bread', 'Butter'],
    ['Bread', 'Butter'],
    ['Milk', 'Bread', 'Butter'],
    ['Milk', 'Bread'],
    ['Bread', 'Butter'],
    ['Milk', 'Butter']
]
```

```
te = TransactionEncoder()
te_data = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_data, columns=te.columns_)
df
```

```

frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)
frequent_itemsets
```

```
rules = association_rules(frequent_itemsets, metric='confidence',
min_threshold=0.6)
rules
```

```
frequent_itemsets['itemsets'] = frequent_itemsets['itemsets'].apply(lambda x: ',
'.join(list(x)))
plt.figure(figsize=(10,5))
plt.bar(frequent_itemsets['itemsets'], frequent_itemsets['support'],
color='skyblue')
plt.xlabel('Itemsets')
plt.ylabel('Support')
plt.title('Support of Frequent Itemsets')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```