Association Rule Mining

→ Also known as Market Basket Analysis

→ We will use Apriori algorithm.

Association rule mining is a technique used to discover relationships between items in large dataset (e.g., finding that customers who by bread often buy milk).

Terminologies

1. Itemset : Collection of items. Ex : {milk, bread}
2. Transaction : Set of items bought together.
3. Association Rule : Generating rules like - If a customer buys X, he will also buy Y. The probability is P%.

Rule Strength Measures

⊛ Support : How frequently the itemset appears in the dataset.

$$\text{Support} = \frac{\text{num transactions containing X and Y}}{\text{Total num of transactions}}$$

⊛ Confidence: How often the rule is true.

$$\text{Confidence} = \frac{\text{Support of X and Y}}{\text{Support of X}}$$

Apriori Algorithm

Step 1: Find frequent itemsets using a level-wise approach that meet the minimum support.

Step 2: Generate rules ~~from~~ from frequent itemsets that meet minimum confidence.

Example: Dataset is given. Find association rules.

TID	Items Purchased
T1	{Bread, Milk}
T2	{Bread, Diapers, Beer, Eggs}
T3	{Milk, Diapers, Beer, Cola}
T4	{Bread, Milk, Diapers, Beer}
T5	{Bread, Milk, Diapers Cola}

Note that, minsup = 40% and minconf = 60%

Solution: Step 01: Find frequent itemsets.

1-itemsets (C_1)

Itemset	Support count	Support (%)
{Bread}	4	4/5 = 80%
{Milk}	4	80%
{Diapers}	4	80%
{Beer}	3	60%
{Eggs}	1	20% ← pruned
{Cola}	2	40%

Frequent 1-itemsets (F_1): {Bread}, {Milk}, {Diapers}, {Beer}, {Cola}

Step 02: 2-itemsets (C_2)

Itemset	Support count	Support (%)	
{Bread, Milk}	3	60%	
{Brad, Diapers}	3	60%	
{Bread, Beer}	2	40%	
{Bread, Cola}	1	20%	← Pruned
{Milk, Diapers}	3	60%	
{Milk, Beer}	2	40%	
{Milk, Cola}	2	40%	
{Diapers, Beer}	3	60%	
{Diapers, Cola}	2	40%	
{Beer, Cola}	1	20%	← Pruned

Frequent 2-itemsets (F_2): {Bread, Milk}, {Bread, Diapers}, {Bread, Beer}, {Milk, diapers}, {Milk, beer}, {Milk, Cola}, {Diapers, beer}, {Diapers, Cola}

3-itemsets (C_3)

Itemset	Support count	Support (%)	
{Bread, Milk, Diapers}	2	40%	
{Bread, Milk, Beer}	1	20%	← prune
{Bread, Diapers, Beer}	2	40%	
{Milk, diapers, Beer}	2	40%	
{Milk, diapers, Cola}	1	20%	← prune

Frequent 3-itemsets (F_3): {Bread, Milk, Diapers}, {Bread, Diapers, Beer}, {Milk, Diapers, Beer}

4-itemsets (C_4)

Itemset
{Bread, Milk - Diapers, Beer}
{ }

Final Step: Generating association rule.

1. {Bread, Milk} → {Diapers}

$$\text{confidence} = \frac{\text{Support}(\{\text{Bread, Milk, Diapers}\})}{\text{Support}(\{\text{Bread, Milk}\})}$$

$$= \frac{2}{3} = 66.7\% > 60\% \quad \checkmark$$

2. {Bread, Diapers} → {Milk}

$$\text{confidence} = \frac{\text{Support}(\{\text{Bread, Diapers, Milk}\})}{\text{Support}(\{\text{Bread, Diapers}\})}$$

$$= \frac{2}{3} = 66.7\% > 60 \quad \checkmark$$

Thus, we can generate more rules.

Interpretation

Rule: {Bread, Milk} → {Diapers}

→ customers who buy bread and milk are 66.7% likely to also buy diapers.

→ This occurs in 40% transactions.