Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

1. Basic Terminology

- **Itemset**: A collection of one or more items. For example, {bread, milk} is an itemset.
- **Support**: The frequency with which an itemset appears in the dataset. If an itemset appears in 10% of transactions, its support is 10%.
- **Confidence**: Measures the likelihood of seeing an itemset B in transactions containing another itemset A. It's calculated as:

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

- **Lift**: Indicates the strength of a rule compared to the random chance of seeing B in transactions. A lift greater than 1 suggests a strong association between A and B.

2. Association Rules

An **association rule** is of the form $A \Rightarrow B$, meaning that if A (the antecedent) is present in a transaction, then B (the consequent) is likely to be present as well.

For example, a rule might be {diaper} $\Rightarrow$ {beer}, suggesting that customers who buy diapers are likely to also buy beer.

3. Mining Process

1. **Step 1**: Find all itemsets with a support above a minimum threshold. This step identifies frequently bought itemsets.
2. **Step 2**: Generate association rules from these frequent itemsets that meet minimum confidence requirements.

4. Algorithms

- **Apriori Algorithm**: Generates frequent itemsets by iteratively increasing the itemset size and pruning infrequent itemsets, making it memory efficient.
- **FP-Growth**: Builds a compact data structure (FP-tree) to generate frequent itemsets without candidate generation, often faster than Apriori.

5. Applications

- **Market Basket Analysis**: Understand customer purchasing patterns.
- **Recommendation Systems**: Suggest products based on commonly associated items.
- **Fraud Detection**: Identify patterns indicative of fraudulent behavior.

2. Association Rules

An **association rule** is of the form $A \Rightarrow B$, meaning that if A (the antecedent) is present in a transaction, then B (the consequent) is likely to be present as well.

For example, a rule might be {diaper} $\Rightarrow$ {beer}, suggesting that customers who buy diapers are likely to also buy beer.

3. Example Calculation with Food Items

Suppose we have the following dataset of transactions in a grocery store:

Transaction ID	Items Purchased
1	{bread, milk, butter}
2	{bread, milk}
3	{bread, butter}
4	{milk, butter}
5	{bread, milk, butter}

We want to analyze the association rule: $\{bread, milk\} \Rightarrow \{butter\}$.

- **Support of** $\{bread, milk, butter\}$: This itemset appears in 2 out of 5 transactions, so:

$$\text{Support}(\{bread, milk, butter\}) = \frac{2}{5} = 0.4$$

- **Support of** $\{bread, milk\}$: This itemset appears in 3 out of 5 transactions, so:

$$\text{Support}(\{bread, milk\}) = \frac{3}{5} = 0.6$$

- **Confidence of** $\{bread, milk\} \Rightarrow \{butter\}$: Calculated as:

$$\text{Confidence}(\{bread, milk\} \Rightarrow \{butter\}) = \frac{\text{Support}(\{bread, milk, butter\})}{\text{Support}(\{bread, milk\})} = \frac{0.4}{0.6} = 0.67$$

- **Lift of** $\{bread, milk\} \Rightarrow \{butter\}$: Suppose $\text{Support}(\{butter\}) = 0.6$. Then,

$$\text{Lift}(\{bread, milk\} \Rightarrow \{butter\}) = \frac{\text{Confidence}(\{bread, milk\} \Rightarrow \{butter\})}{\text{Support}(\{butter\})} = \frac{0.67}{0.6} \approx 1.12$$

Since the lift is greater than 1, this suggests a positive association between {bread, milk} and {butter}.

4. Mining Process

1. **Step 1**: Find all itemsets with a support above a minimum threshold. This step identifies frequently bought itemsets.
2. **Step 2**: Generate association rules from these frequent itemsets that meet minimum confidence requirements.

Apriori Algorithm

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

Introduction

The Apriori Algorithm is a classic algorithm in Association Rule Mining, used to find frequent itemsets in a dataset and derive association rules. The key idea is that any subset of a frequent itemset must also be frequent, allowing the algorithm to reduce the number of itemsets examined and improve efficiency.

Steps of the Apriori Algorithm

1. **Set a Minimum Support and Confidence Threshold**: Choose a minimum support threshold to identify frequent itemsets and a minimum confidence threshold for the rules.
2. **Find Frequent Itemsets**: Identify all itemsets that meet the minimum support threshold. Start with 1-itemsets, then expand to larger itemsets by combining items that meet the minimum support requirement.
3. **Generate Association Rules**: For each frequent itemset, generate possible association rules. Keep rules that meet the minimum confidence threshold.

Example with a Food Dataset

Suppose we have the following transactions in a grocery store:

Transaction ID	Items Bought
T1	Bread, Milk
T2	Bread, Diaper, Beer
T3	Milk, Diaper, Beer, Eggs
T4	Bread, Milk, Diaper, Beer
T5	Bread, Milk, Diaper

Let the minimum support threshold be 60% (0.6), and the minimum confidence threshold be 70% (0.7).

Step 1: Generate 1-Itemsets and Calculate Support

We calculate the support for each individual item:

- **Bread**: 4 transactions (T1, T2, T4, T5) $\rightarrow$ Support $= \frac{4}{5} = 0.8$
- **Milk**: 4 transactions (T1, T3, T4, T5) $\rightarrow$ Support $= \frac{4}{5} = 0.8$
- **Diaper**: 4 transactions (T2, T3, T4, T5) $\rightarrow$ Support $= \frac{4}{5} = 0.8$
- **Beer**: 3 transactions (T2, T3, T4) $\rightarrow$ Support $= \frac{3}{5} = 0.6$
- **Eggs**: 1 transaction (T3) $\rightarrow$ Support $= \frac{1}{5} = 0.2$

Only items meeting the minimum support of 0.6 are considered for the next step. Thus, **Eggs** is discarded.

Step 2: Generate 2-Itemsets and Calculate Support

Next, we create pairs from the remaining items and calculate their support:

- {Bread, Milk}: 3 transactions (T1, T4, T5) → Support = $\frac{3}{5} = 0.6$
- {Bread, Diaper}: 3 transactions (T2, T4, T5) → Support = $\frac{3}{5} = 0.6$
- {Bread, Beer}: 2 transactions (T2, T4) → Support = $\frac{2}{5} = 0.4$ (discarded)
- {Milk, Diaper}: 3 transactions (T3, T4, T5) → Support = $\frac{3}{5} = 0.6$
- {Milk, Beer}: 2 transactions (T3, T4) → Support = $\frac{2}{5} = 0.4$ (discarded)
- {Diaper, Beer}: 3 transactions (T2, T3, T4) → Support = $\frac{3}{5} = 0.6$

Only itemsets meeting the minimum support of 0.6 are retained. We discard {Bread, Beer} and {Milk, Beer}.

Step 3: Generate 3-Itemsets and Calculate Support

We combine the remaining 2-itemsets to form 3-itemsets:

- {Bread, Milk, Diaper}: 2 transactions (T4, T5) → Support = $\frac{2}{5} = 0.4$ (discarded)
- {Bread, Diaper, Beer}: 2 transactions (T2, T4) → Support = $\frac{2}{5} = 0.4$ (discarded)
- {Milk, Diaper, Beer}: 2 transactions (T3, T4) → Support = $\frac{2}{5} = 0.4$ (discarded)

No 3-itemsets meet the minimum support threshold, so we stop here.

Step 4: Generate Association Rules

Using the frequent 2-itemsets, we generate association rules. For example, from {Bread, Milk}:

1. **Bread → Milk**:
   Confidence = $\frac{\text{Support(Bread and Milk)}}{\text{Support(Bread)}} = \frac{0.6}{0.8} = 0.75$ (meets confidence threshold)
2. **Milk → Bread**:
   Confidence = $\frac{\text{Support(Bread and Milk)}}{\text{Support(Milk)}} = \frac{0.6}{0.8} = 0.75$ (meets confidence threshold)

Summary of Results

After applying the Apriori Algorithm, the frequent itemsets are:

- **1-itemsets**: Bread, Milk, Diaper, Beer
- **2-itemsets**: {Bread, Milk}, {Bread, Diaper}, {Milk, Diaper}, {Diaper, Beer}

The valid association rules are:

- **Bread → Milk** (Support: 0.6, Confidence: 0.75)
- **Milk → Bread** (Support: 0.6, Confidence: 0.75)

These rules can help in strategies like product placement and promotional bundles in the store.