ARITIFICAL INTELLIGENCE

PHASE-4 SUBMISSION

Market Basket Insights

Market basket analysis is a valuable technique in retail and e-commerce that aims to uncover patterns and associations among items frequently purchased together. Understanding these item associations allows businesses to optimize their marketing strategies, cross-sell products, and enhance the overall shopping experience. The Apriori algorithm is a widely used tool for conducting market basket analysis.

What is the Apriori Algorithm?

The Apriori algorithm is a classic data mining technique used to discover frequent itemsets and generate association rules in large transactional datasets. It was introduced by Rakesh Agrawal and Ramakrishnan Srikant in 1994 and has since become a fundamental tool for finding patterns in retail and other domains.

The key idea behind Apriori is that if an itemset is frequent (i.e., it occurs often enough) in a transaction database, then all of its subsets must also be frequent. By iteratively identifying frequent itemsets, the algorithm can generate association rules that provide insights into which items are commonly bought together.

How Does the Apriori Algorithm Work?

The Apriori algorithm works in a series of steps:

1. Itemset Generation: It begins by generating a list of all individual items present in the dataset. These are called "1-itemsets."
2. Finding Frequent 1-Itemsets: The algorithm scans the dataset to find the support (occurrence frequency) of each 1-itemset. Those with support exceeding a predefined threshold are deemed frequent itemsets.
3. Combining Itemsets: The algorithm then generates 2-itemsets by combining pairs of frequent 1-itemsets. It proceeds to find the support of these 2-itemsets.
4. Recursive Process: This process continues, generating 3-itemsets, 4-itemsets, and so on, until no more frequent itemsets can be found.

5. Association Rules: Once all frequent itemsets are determined, Apriori generates association rules. An association rule consists of an antecedent (the left-hand side) and a consequent (the right-hand side). For example, "If a customer buys A and B, they are likely to buy C." The strength of these rules is measured by metrics like confidence and lift.

Usage and Applications:

The Apriori algorithm is commonly used in the following areas:

1. Market Basket Analysis: Retailers and e-commerce platforms use Apriori to analyse customers' purchase histories, identify items that are frequently bought together, and optimize product placements.
2. Cross-Selling: Apriori helps businesses identify opportunities for cross-selling additional items based on what customers are already purchasing.
3. Inventory Management: Retailers use Apriori to make informed decisions about stock levels, ensuring popular items are in sufficient supply.
4. Recommendation Systems: E-commerce websites and streaming services use Apriori's association rules to make product or content recommendations to users.
5. Customer Segmentation: By understanding item associations, businesses can segment customers and tailor marketing efforts accordingly.

Comparison between Apriori Algorithm and Other Algorithms:

1. **FP-Growth (Frequent Pattern Growth):**
   Reason to Choose Apriori: Apriori is chosen when transparency and interpretability are important. The FP-Growth algorithm can be more efficient but is less intuitive when it comes to explaining the results
2. **Eclat (Equivalence Class Transformation):**
   Reason to Choose Apriori: Apriori is often favored due to its interpretability. Eclat focuses on vertical data format, which might be less intuitive to non-technical stakeholders.
3. **Association Rule Learning with Machine Learning Models (e.g., Random Forest, XGBoost):**
   Reason to Choose Apriori: Apriori is selected when the focus is on item association discovery rather than predictive modeling. Machine learning models can provide predictions but lack the transparency and simplicity of Apriori.

4. **Collaborative Filtering (User-Item Interaction-Based Recommendations):**
   Reason to Choose Apriori: Apriori is preferred when understanding item associations without considering individual user behaviour is the goal. Collaborative filtering focuses on personalized recommendations, which might not be the primary objective.
5. **Advanced Metric Analysis with Apriori:**
   Reason to Choose Apriori: Apriori is suitable when straightforward, actionable insights are needed. While advanced metrics can enhance Apriori's rules, Apriori remains a user-friendly choice.
6. **Sequential Pattern Mining (SPM):**
   Reason to Choose Apriori: Apriori is used when traditional association rules based on itemsets are required. SPM deals with sequences of events or transactions, which might not be necessary for market basket analysis.
7. **Deep Learning Models (e.g., Recurrent Neural Networks):**
   Reason to Choose Apriori: Apriori is selected when item associations based on purchase patterns are the main focus. Deep learning models require more data and are typically used for complex, large-scale recommendations.
8. **Clustering Techniques (e.g., K-Means, DBSCAN):**
   Reason to Choose Apriori: Apriori is used when discovering item associations is the primary goal. Clustering techniques group similar items or users, which may not be as informative for direct market basket insights.

Python Model For Market Basket Insights:

```
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

# Load your dataset (replace 'your_dataset.csv' with the actual dataset file path)
data = pd.read_csv('Assignment-1_Data.csv', sep=';')

# Select relevant columns ('BillNo' and 'Itemname')
data = data[['BillNo', 'Itemname']]
```

```
# Convert the data into a one-hot encoded format
basket = (data.groupby(['BillNo', 'Itemname'])['Itemname']
        .count().unstack().reset_index().fillna(0)
        .set_index('BillNo'))

# Convert item counts to 1 or 0
basket_sets = basket.applymap(lambda x: 1 if x > 0 else 0)
# Convert item counts to boolean (True/False) values
basket_sets = basket.applymap(lambda x: x > 0).astype(bool)

frequent_itemsets = apriori(basket_sets, min_support=0.01,
use_colnames=True)

# Generate association rules with a lower confidence threshold
association_rules = association_rules(frequent_itemsets, metric="lift",
min_threshold=0.01)

# Filter and interpret the rules based on your specific requirements

# Print the frequent item sets
print("Frequent Item Sets:")
print(frequent_itemsets)

# Print the association rules
print("\nAssociation Rules:")
print(association_rules)
```

Output:

Frequent Item Sets:

	support	itemsets
0	**0.017248**	**(10 COLOUR SPACEBOY PEN)**
1	**0.012094**	**(12 IVORY ROSE PEG PLACE SETTINGS)**
2	**0.013085**	**(12 MESSAGE CARDS WITH ENVELOPES)**
3	**0.017645**	**(12 PENCIL SMALL TUBE WOODLAND)**
4	**0.027359**	**(12 PENCILS SMALL TUBE RED RETROSPOT)**
...	**...**	**...**
2859	**0.010508**	**(JUMBO BAG RED RETROSPOT, JUMBO BAG BAROQUE B...**

```
2860  0.010111  (JUMBO BAG OWLS, JUMBO BAG RED RETROSPOT, JUMB...
2861  0.010508  (JUMBO BAG WOODLAND ANIMALS, JUMBO BAG OWLS, J...
2862  0.010508  (JUMBO BAG OWLS, JUMBO BAG RED RETROSPOT, JUMB...
2863  0.010309  (POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...

[2864 rows x 2 columns]

Association Rules:
                          antecedents  \
0          (12 PENCILS SMALL TUBE RED RETROSPOT)
1               (12 PENCILS SMALL TUBE SKULL)
2           (3 PIECE SPACEBOY COOKIE CUTTER SET)
3             (GINGERBREAD MAN COOKIE CUTTER)
4               (FELTCRAFT 6 FLOWER FRIENDS)
...                              ...
5935  (POPPY'S PLAYHOUSE BATHROOM, POPPY'S PLAYHOUSE...
5936               (POPPY'S PLAYHOUSE KITCHEN)
5937             (POPPY'S PLAYHOUSE LIVINGROOM)
5938              (POPPY'S PLAYHOUSE BATHROOM)
5939               (POPPY'S PLAYHOUSE BEDROOM)

                          consequents  antecedent support  \
0               (12 PENCILS SMALL TUBE SKULL)          0.027359
1          (12 PENCILS SMALL TUBE RED RETROSPOT)          0.022205
2             (GINGERBREAD MAN COOKIE CUTTER)          0.025575
3           (3 PIECE SPACEBOY COOKIE CUTTER SET)          0.045202
4                (3 STRIPEY MICE FELTCRAFT)          0.039056
...                              ...           ...
5935  (POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...          0.012292
5936  (POPPY'S PLAYHOUSE LIVINGROOM, POPPY'S PLAYHOU...
0.028945
```

5937	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...	0.022403
5938	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...	0.014869
5939	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...	0.028549

	consequent support	support	confidence	lift	leverage \
0	0.022205	0.015067	0.550725	24.802277	0.014460
1	0.027359	0.015067	0.678571	24.802277	0.014460
2	0.045202	0.010111	0.395349	8.746226	0.008955
3	0.025575	0.010111	0.223684	8.746226	0.008955
4	0.023394	0.010309	0.263959	11.283145	0.009396
...	...	...	...	...	...
5935	0.016257	0.010309	0.838710	51.590873	0.010109
5936	0.010904	0.010309	0.356164	32.663512	0.009994
5937	0.011499	0.010309	0.460177	40.019530	0.010052
5938	0.014869	0.010309	0.693333	46.628978	0.010088
5939	0.010309	0.010309	0.361111	35.027778	0.010015

	conviction	zhangs_metric
0	2.176383	0.986676
1	3.025993	0.981474
2	1.579089	0.908910
3	1.255192	0.927594
4	1.326837	0.948414
...	...	...
5935	6.099207	0.992820
5936	1.536255	0.998280
5937	1.831158	0.997356
5938	3.212383	0.993324
5939	1.549081	1.000000

[5940 rows x 10 columns]

Interpretation of The Output:

The output contains a list of frequent item sets and association rules discovered in your market basket data. Analysts use these rules to identify item associations and make business decisions. For example, a rule with high confidence and lift might suggest placing related items together on store shelves, while a low-confidence rule might not provide actionable insights.

Frequent Item Sets:

support: This column indicates the support of each item set. Support is the proportion of transactions in the dataset that contain the item set. For example, the item set (10 COLOUR SPACEBOY PEN) has a support of 0.017248, meaning it appears in approximately 1.72% of the transactions.

itemsets: This column lists the items or itemset that are considered frequent, enclosed in parentheses. Each row represents a unique frequent itemset along with its corresponding support.

Association Rules:

antecedents: This column displays the antecedent item(s) of the association rule, i.e., the items on the left-hand side of the rule.

consequents: This column displays the consequent item(s) of the association rule, i.e., the items on the right-hand side of the rule.

antecedent support: The support of the antecedent item(s), indicating how often the antecedent(s) appear in transactions.

consequent support: The support of the consequent item(s), indicating how often the consequent(s) appear in transactions.

support: This column represents the support for the combined antecedent and consequent item(s). It shows how frequently the whole rule appears in transactions.

confidence: Confidence measures the reliability of the association rule. It's the proportion of transactions that contain the consequent item(s) given that they contain the antecedent item(s). For instance, a confidence of 0.550725 indicates that the rule is correct in approximately 55.07% of cases.

lift: Lift indicates how much more likely the consequent item(s) are purchased when the antecedent item(s) are also purchased compared to when the purchase of the consequent is independent of the antecedent. Lift greater than 1 suggests a positive association.

leverage: Leverage quantifies the difference in the support between the antecedent and consequent appearing together in transactions and what would be expected if they were independent. Positive values indicate that the rule's antecedent and consequent are purchased together more often than expected.

conviction: Conviction measures the implication strength of the rule. It indicates how much more likely the rule's consequent is bought when the rule's antecedent is true compared to when it is false. Higher values suggest stronger associations.

zhangs_metric: This is Zhang's metric, which is another measure of rule interest. It varies between 0 and 1, with higher values indicating stronger associations.

TEAM MEMBERS:

1. J. ARUN KUMAR-813821205002
2. C. MURUGANANTHAN-813821205033
3. T. SRIHARIHARAN-813821205049
4. S. VENGADASHAN-813821205055
5. S. VISHWA-813821205059