MIT | Academy of Engineering
(An Autonomous Institute Affiliated to Savitribai Phule University)

Apriori Algorithm

Develop apriori algorithm for finding frequent itemsets and suggest association rules for Bakery.

Name : Aditya Joshi
PRN No : 202101040132
Roll No : 104
Batch : AIL3

INDEX

ML PROJECT INDEX

01 Introduction

02 Problem Statement

03 Association Rule

04 Apriori Algorithm

05 Results

06 Conclusion

Introduction

In this project, we delve into the complexities of the Apriori algorithm, an influential method in the realm of machine learning for uncovering frequent itemsets in extensive datasets. Our ambition goes beyond mere detection of patterns; we aspire to harness the potential of this algorithm to unearth significant associations and formulate astute rules, especially designed for the unique environment of bakeries.

Introduction to Apriori Algorithm

Apriori is the most famous frequent pattern mining method. It scans dataset repeatedly and generate item sets by bottom-top approach.

Apriori algorithm is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties

Dataset

The dataset belongs to "The Bread Basket" a bakery located in Edinburgh. The dataset provide the transaction details of customers who ordered different items from this bakery online during the time period from 26-01-11 to 27-12-03. The dataset has 20507 entries, over 9000 transactions, and 4 columns.

- **TransactionNo** : unique identifier for every single transaction
- Items : items purchased
- **DateTime** : date and time stamp of the transactions
- **Daypart** : part of the day when a transaction is made (morning, afternoon, evening, night)
- **DayType** : classifies whether a transaction has been made in weekend or weekdays

Association Rules

Association rule learning is a kind of unsupervised learning technique that tests for the reliance of onedata element on another data element and design appropriately so that it can be more cost-effective.

THE GIVEN THREE COMPONENTS COMPRISE THE APRIORI ALGORITHM.

Support

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)
= 400/4000 = 10 percent.

THE GIVEN THREE COMPONENTS COMPRISE THE APRIORI ALGORITHM.

Confidence

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

Hence,

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

THE GIVEN THREE COMPONENTS COMPRISE THE APRIORI ALGORITHM.

Lift

Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.
Lift = (Confidence (Biscuits – chocolates)/ (Support (Biscuits)

= 50/10 = 5
It means that the probability of people buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone. If the lift value is below one, it requires that the people are unlikely to buy both the items together. Larger the value, the better is the combination.

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

Step 1

Determine the support of itemsets in thetransactional database, and select the minimumsupport and confidence

Step 2

Take all supports in the transaction with highersupport value than the minimum or selected supportvalue.

Step 3

Find all the rules of these subsets that have higherconfidence value than the threshold or minimumconfidence.

Step 4

Sort the rules as the decreasing order of lift.

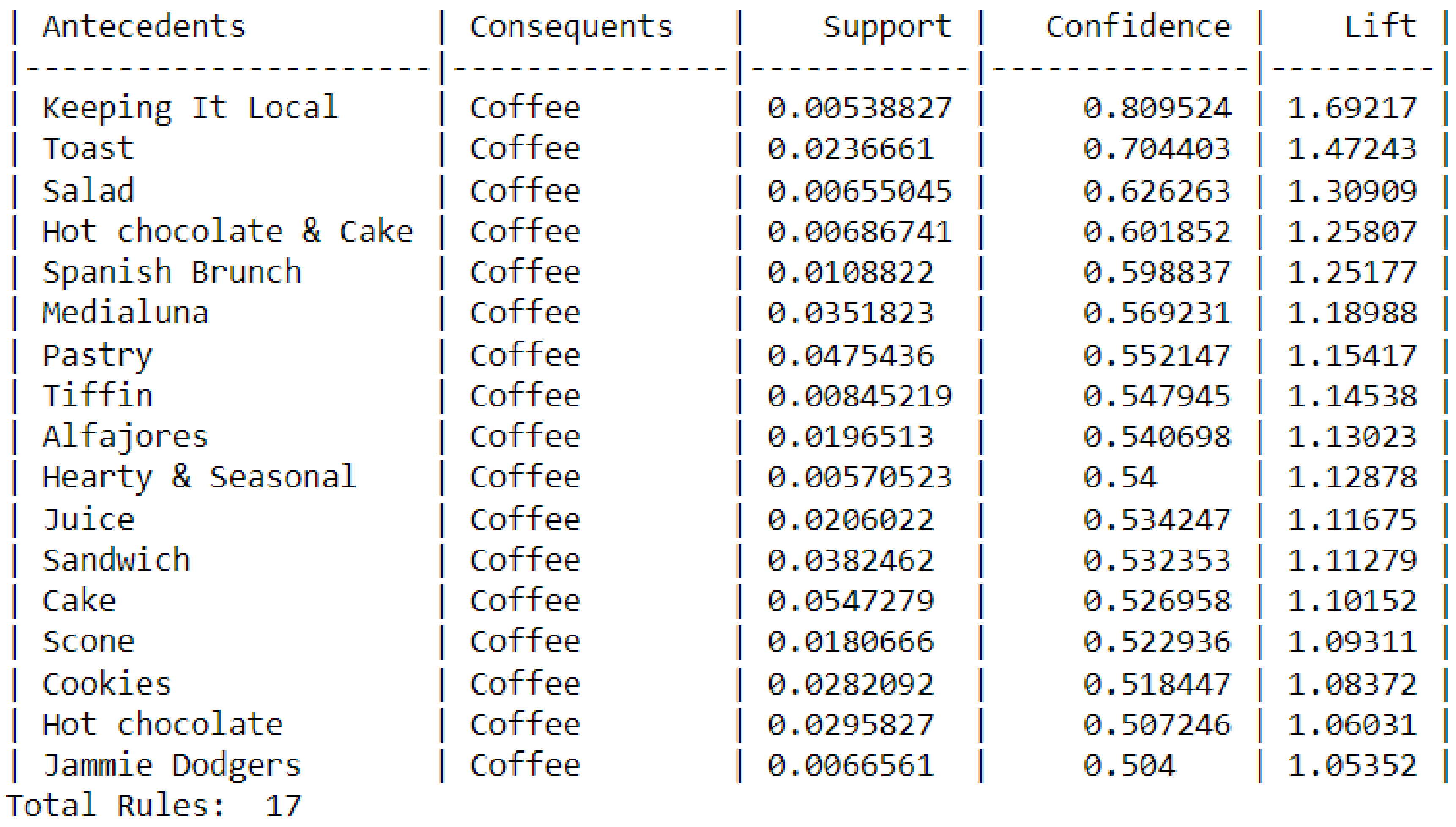
Advantages

- This is easy to understand algorithm
- The join and prune steps of the algorithm can be easily implemented on large datasets.

Disadvantages

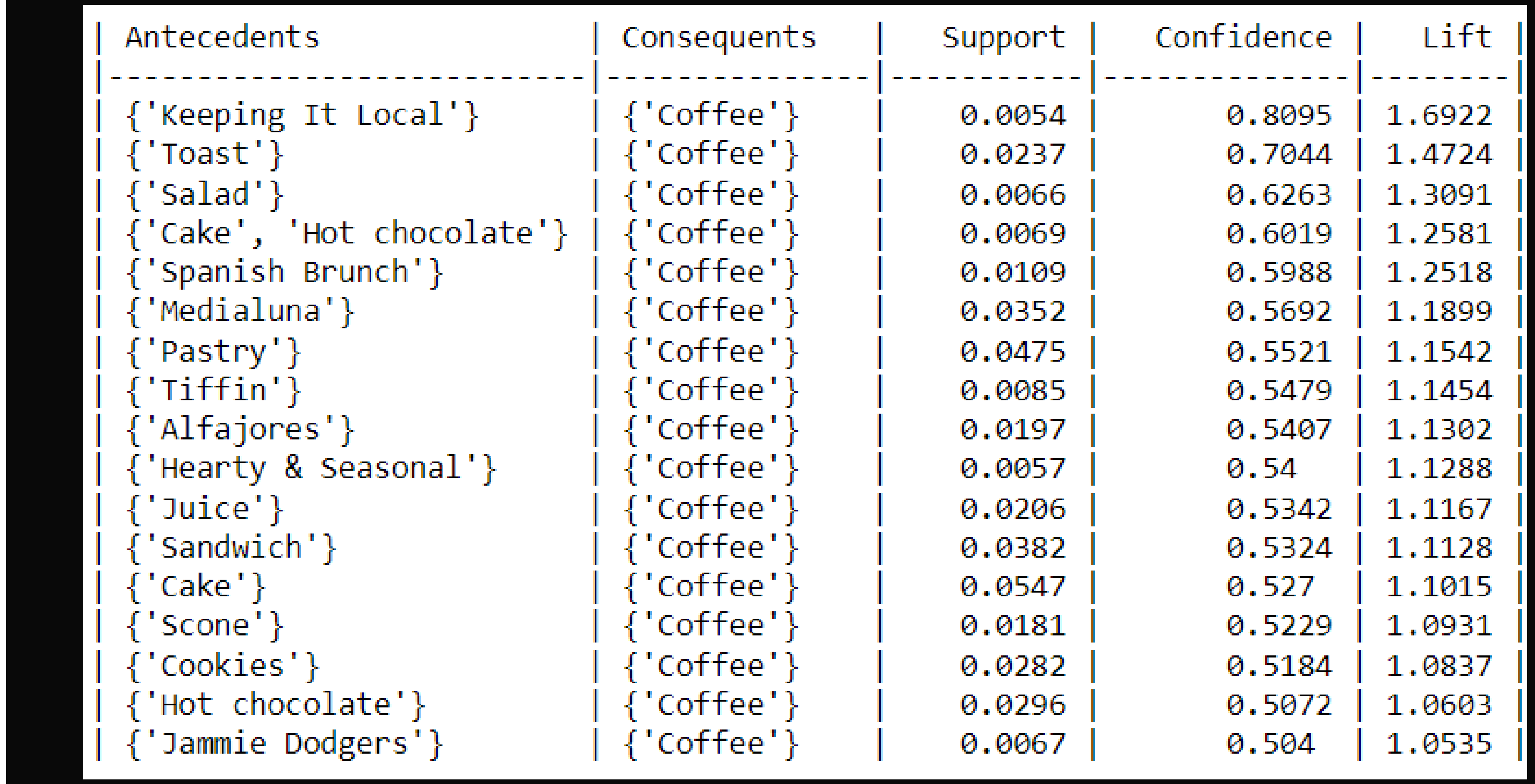
- The apriori algorithm works slow compared to other algorithms.
- The overall performance can be reduced as it scans the database for multiple times.
- The time complexity and space complexity of the apriori algorithm is O(2D), which is very high. Here D represents the horizontal width present in the database.

With Libraries



```
| Antecedents           | Consequents    |    Support |   Confidence |    Lift |
|-----------------------|----------------|------------|--------------|---------|
| Keeping It Local      | Coffee         | 0.00538827 |     0.809524 | 1.69217 |
| Toast                 | Coffee         | 0.0236661  |     0.704403 | 1.47243 |
| Salad                 | Coffee         | 0.00655045 |     0.626263 | 1.30909 |
| Hot chocolate & Cake  | Coffee         | 0.00686741 |     0.601852 | 1.25807 |
| Spanish Brunch        | Coffee         | 0.0108822  |     0.598837 | 1.25177 |
| Medialuna             | Coffee         | 0.0351823  |     0.569231 | 1.18988 |
| Pastry                | Coffee         | 0.0475436  |     0.552147 | 1.15417 |
| Tiffin                | Coffee         | 0.00845219 |     0.547945 | 1.14538 |
| Alfajores             | Coffee         | 0.0196513  |     0.540698 | 1.13023 |
| Hearty & Seasonal     | Coffee         | 0.00570523 |     0.54     | 1.12878 |
| Juice                 | Coffee         | 0.0206022  |     0.534247 | 1.11675 |
| Sandwich              | Coffee         | 0.0382462  |     0.532353 | 1.11279 |
| Cake                  | Coffee         | 0.0547279  |     0.526958 | 1.10152 |
| Scone                 | Coffee         | 0.0180666  |     0.522936 | 1.09311 |
| Cookies               | Coffee         | 0.0282092  |     0.518447 | 1.08372 |
| Hot chocolate         | Coffee         | 0.0295827  |     0.507246 | 1.06031 |
| Jammie Dodgers        | Coffee         | 0.0066561  |     0.504    | 1.05352 |
Total Rules:  17
```

Without Libraries



```
| Antecedents                | Consequents   |   Support |   Confidence |   Lift |
|----------------------------|---------------|-----------|--------------|--------|
| {'Keeping It Local'}       | {'Coffee'}    |    0.0054 |       0.8095 | 1.6922 |
| {'Toast'}                  | {'Coffee'}    |    0.0237 |       0.7044 | 1.4724 |
| {'Salad'}                  | {'Coffee'}    |    0.0066 |       0.6263 | 1.3091 |
| {'Cake', 'Hot chocolate'}  | {'Coffee'}    |    0.0069 |       0.6019 | 1.2581 |
| {'Spanish Brunch'}         | {'Coffee'}    |    0.0109 |       0.5988 | 1.2518 |
| {'Medialuna'}              | {'Coffee'}    |    0.0352 |       0.5692 | 1.1899 |
| {'Pastry'}                 | {'Coffee'}    |    0.0475 |       0.5521 | 1.1542 |
| {'Tiffin'}                 | {'Coffee'}    |    0.0085 |       0.5479 | 1.1454 |
| {'Alfajores'}              | {'Coffee'}    |    0.0197 |       0.5407 | 1.1302 |
| {'Hearty & Seasonal'}      | {'Coffee'}    |    0.0057 |       0.54   | 1.1288 |
| {'Juice'}                  | {'Coffee'}    |    0.0206 |       0.5342 | 1.1167 |
| {'Sandwich'}               | {'Coffee'}    |    0.0382 |       0.5324 | 1.1128 |
| {'Cake'}                   | {'Coffee'}    |    0.0547 |       0.527  | 1.1015 |
| {'Scone'}                  | {'Coffee'}    |    0.0181 |       0.5229 | 1.0931 |
| {'Cookies'}                | {'Coffee'}    |    0.0282 |       0.5184 | 1.0837 |
| {'Hot chocolate'}          | {'Coffee'}    |    0.0296 |       0.5072 | 1.0603 |
| {'Jammie Dodgers'}         | {'Coffee'}    |    0.0067 |       0.504  | 1.0535 |
```

Conclusion

01 Online classrooms mean digital learning for everyone.

02 A global market for practical courses and credentials.

03 Improvement in the quality of blended learning

04 Rising demand for skills-based programs.

05 Greater investment on interactive technology in solving the digital divide