CIND110 Team Assignment Role 3: Data Mining Analyst

By Stephanie Boissonneault

Role Description:

- Employ data mining techniques to discover patterns and correlations among products.
- Key role in identifying trends and insights for business decisions.

GroceryTransactions

TransID	item
1	Bread, Milk
2	Bread, Cookies, Juice, Eggs
3	Milk, Cookies, Juice, Soda
4	Bread, Milk, Juice, Eggs
5	Bread, Milk, Eggs, Soda
6	Bread, Milk, Juice, Coffee
7	Bread, Milk, Eggs
8	Coffee, Cookies
9	Bread, Coffee, Eggs
10	Milk, Eggs

Tasks:

1. **Use a 0.2 minimum support value to apply the apriori algorithm on the attached grocery store transaction dataset, detailing the process and identifying frequent itemsets.**

The apriori algorithm uses downward closure and antimontonicity to help reduce the search

Before proceeding to apply the apriori algorithm, I will first provide a brief description and details of key terms that will be used during the process:

- Itemset is the set of items purchased by customers (LHS ∪ RHS).
- Support is the frequency at which a specific itemset appears in the collection of transactions, usually described as a percentage. It is the count of transactions in which the itemsets (LHS ∪ RHS) appear divided by m transactions (number of transactions).

When applying the apriori algorithm, we will be considering a minimum support value of 0.2 as outlined in the question. In other words, itemsets with a resulting support value lower than 0.2 will be filtered out during the process to help filter out itemsets that have less overwhelming evidence that the items in the itemset occur together.

#1. Computing support for each one-item itemsets by scanning the database once and counting the number of transactions that the item ij appears in and add to *Candicate frequent one-item "C1"* table.

Candidate frequent one-itemset "C1"		
Itemset [LHS ∪ RHS]	**Support Calculation:** count [LHS ∪ RHS] / m transactions or Support (ij) = count(ij)/m	**Support**
Bread	7/10	0.7
Milk	7/10	0.7
Cookies	3/10	0.3
Juice	4/10	0.3
Eggs	6/10	0.6
Soda	2/10	0.2
Coffee	3/10	0.3

#2. In this first example, all items qualify for the *Frequent one-item itemset "L1"* table since the support value of each one-item itemset is equal to above the threshold as set by the 0.2 minimum support value. In other words, all (ij) support are >= 0.2 min and becomes the frequent 1-itemset "L1".

Frequent one-item itemsets “L1”	
Itemset	**Support**
Bread	0.7
Milk	0.7
Cookies	0.3
Juice	0.3
Eggs	0.6
Soda	0.2
Coffee	0.3

#3. Creating (k + 1)-itemsets (a two-item “itemset”) by combining common members from “L1”.

Candidate frequent two-itemset “C2”		
Itemset [LHS ∪ RHS]	**Support calculations** count [LHS ∪ RHS] / m transactions	**Support**
{bread, milk}	5/10	0.5
{bread, cookies}	1/10	0.1
{bread, juice}	3/10	0.3
{bread, eggs}	5/10	0.5
{bread, soda}	1/10	0.1
{bread, coffee}	2/10	0.2
{milk, cookies}	1/10	0.1
{milk, juice}	3/10	0.3
{milk, eggs}	4/10	0.4
{milk, soda}	2/10	0.2
{milk, coffee}	1/10	0.1
{cookies, juice}	2/10	0.2

{cookies, eggs}	1/10	0.1
{cookies, soda}	1/10	0.1
{cookies, coffee}	1/10	0.1
{juice, eggs}	2/10	0.2
{juice, soda}	1/10	0.1
{juice, coffee}	1/10	0.1
{eggs, soda}	1/10	0.1
{eggs, coffee}	0/10	0.0
{soda, coffee}	0/10	0.0

#4. Only C2 itemsets that are equal to or over the 0.2 minimum support value threshold are selected for the *Frequent two-itemsets "L2"* table as listed in the table below:

Frequent two-itemsets "L2"	
Itemset	**Support**
{bread, milk}	0.5
{bread, juice}	0.3
{bread, eggs}	0.5
{bread, coffee}	0.2
{milk, juice}	0.3
{milk, eggs}	0.4
{milk, soda}	0.2
{cookies, juice}	0.2
{juice, eggs}	0.2

#5. Next, I create a candidate a *Frequent three-item itemsets "C3"* table where all items in the three-item itemsets are contained as two-item subsets in "L2". The subset {bread, juice, cookies} for example could not exist in the candidate three-item itemset because although {bread, juice} and {juice, cookies} are subsets of L2, the two-item itemset {bread, cookies} does not exist in "L2".

Candidate frequent three-item itemset “C3”		
Itemset [LHS ∪ RHS]	**Support calculations** count [LHS ∪ RHS] / m transactions	**Support**
{bread, milk, juice}	2/10	0.2
{bread, milk, eggs}	2/10	0.2
{bread, juice, eggs}	2/10	0.2
{milk, juice, eggs}	1/10	0.1

Note that all three-item itemsets except for the {milk, juice, eggs} itemset are selected for the Frequent three-item itemset “L3” since they have a support value that is >= 0.2 which is the minimum support threshold.

Frequent three-item itemsets “L3”	
Itemset	**Support**
{bread, milk, juice}	0.2
{bread, milk, eggs}	0.3
{bread, juice, eggs}	0.2

2. **Generate rules with at least 0.5 confidence for itemsets of three items, explaining the significance of these findings.**

When looking to generate rules with a confidence of at least 0.5, we are looking to generate rules of a higher strength that have a >= 50% probability that the items in the RHS will indeed be purchased given that the items in the LHS will be purchased.

To answer this question, I will outline all possible rules using three items from the three-item itemsets listed in the Frequent three-item itemsets “L3” table, and calculate the confidence of each rule using the following formula:

Confidence = (support (LHS U RHS) / (support(LHS))

Recall that support is the number of transactions containing the itemset divided by the total number of transactions.

Confidence of Associations Rules for Frequent three-item itemsets from “L3”			
Itemset	**Association rule** [LHS => RHS]	**Confidence calculations** Support (LHS U RHS) / support(LHS)	**Confidence**
{bread, milk, juice}	Bread => milk, juice	(2/10) / (7/10) = 0.2 / 0.7	0.29
	Bread, milk => juice	(2/10) / (5/10) = 0.2 / 0.5	0.4
	Bread, juice => milk	(2/10) / (3/10) = 0.2 / 0.3	0.67
	Milk => bread, juice	(2/10) / (7/10) = 0.2 / 0.7	0.29
	Milk, juice => bread	(2/10) / 3/10) = 0.2 / 0.3	0.67
	Juice => bread, milk	(2/10) / (4/10) = 0.2 / 0.4	0.5
{bread, milk, eggs}	Bread => milk, eggs	(3/10) / 7(10) = 0.3 / 0.7	0.43
	Bread, milk => eggs	(3/10) / (5/10) = 0.3 / 0.5	0.6
	Bread, eggs => milk	(3/10) / (5/10) = 0.3 / 0.5	0.6
	Milk => bread, eggs	(3/10) / (7/10) = 0.3 / 0.7	0.43
	Milk, eggs => bread	(3/10) / (4/10) = 0.3 / 0.4	0.75
	Eggs => bread, milk	(3/10) / (6/10) = 0.3 / 0.6	0.5
{bread, juice, eggs}	Bread => juice, eggs	(2/10) / (7/10) = 0.2 / 0.7	0.29
	Bread, juice => eggs	(2/10) / (3/10) = 0.2 / 0.3	0.67
	Bread, eggs => juice	(2/10) / (5/10) = 0.2 / 0.5	0.4

	Juice => bread, eggs	(2/10) / (4/10) = 0.2 / 0.4	0.5
	Juice, eggs => bread	(2/10) / (2/10) = 0.2 / 0.2	1
	Eggs => bread, juice	(2/10) / (6/10) = 0.2 / 0.6	0.34

Next, I will filter out rules which have a confidence level below 0.5. The following table outlines the association rules that have a confidence interval of at least 0.5 and provides a summary of the meaning of each rule.

Association rule [LHS => RHS]	**Confidence level**	**Association rule explanation**
Bread, juice => milk	0.67	Shoppers who purchase bread and juice have a 67% probability of also purchasing milk.
Milk, juice => bread	0.67	Shoppers who purchase milk and juice have a 67% probability of also purchasing bread.
Juice => bread, milk	0.5	Shoppers who purchase juice have a 50% probability of also purchasing bread and milk.
Bread, milk => eggs	0.6	Shoppers who purchase bread and milk have a 60% probability of also purchasing eggs.
Bread, eggs => milk	0.6	Shoppers who purchase bread and eggs have a 60% probability of also purchasing milk.
Milk, eggs => bread	0.75	Shoppers who purchase milk and eggs have a 75% probability of also purchasing bread.
Eggs => bread, milk	0.5	Shoppers who purchase eggs have a 50% probability of also purchasing bread and milk.
Bread, juice => eggs	0.67	Shoppers who purchase bread and juice have a 67% probability of also purchasing eggs.
Juice => bread, eggs	0.5	Shoppers who purchase juice have a 50% probability of also purchasing bread and eggs.
Juice, eggs => bread	1	Shoppers who purchase juice and eggs have a 100% probability of also purchasing bread.

Rational behind the strengths of these rules:
The strengths of each association rule are associated with the confidence level. An association rule with a stronger confidence level is said to be stronger because we can be more confident that the items from the RHS will be purchased along with items from the LHS since higher

confidence represents a higher probability that the RHS item does have a relationship with the LHS item.

- The association rule Juice, eggs => bread was found to be the strongest out of all the three-item itemset association rules with a confidence level of 1. In other words, the probability of bread being purchased by a customer, given that they purchase Juice and eggs, is 100%.

- There is also a high likelihood (75% probability) that a customer will purchase bread if they've decided to purchase milk and eggs.

Both these examples provide association rules that are quite strong because we can be quite confident that the items from the itemset RHS will be purchased given the LHS is purchased.

Significance of association rules:

The association rules identified in the above table can be used to help the grocery store business predict what items are most likely to be purchased together by customers, and help them make business decisions and implement marketing strategies. By understanding the likelihood of certain grocery items being purchased together, the grocery store might use these insights to better understand and meet customer's wants and needs. They might for example:

- Implement a strategy that ensures that items frequently bought together such as juice, eggs, and bread always have enough inventory stock in stock. If workers notice the inventory for juice and eggs is getting low for example, they should also check bread inventory and order more as well to ensure that all items are available for purchase to customers for their shopping convenience).
- Strategize where items are stored in their store layout to encourage customers to walk through more of the grocery store. For example, placing milk and eggs on the opposite side of the store from the bread to encourage customers to walk through the store and browse more items, knowing very well that if the customer is to purchase milk and eggs, they are highly likely to also walk through the store to purchase bread.
- Alternatively, perhaps the grocery store is small and may choose to put these items closer together in the store to increase convenience for customers.
- Information obtained from understanding what items are frequently bought together could be used to create algorithms for customer loyalty programs to advertise the sale of certain items or the customer's ability to acquire extra points from purchasing a certain item, based on other items they have been known to purchase. For example, have the app tell the customer they can earn extra points on purchasing a specific new expensive bread if they are known to frequently buy milk and eggs, for example.

3. **Discuss why specific rules exhibit strong associations based on confidence and support metrics, using examples from the analysis.**

Association rules are rules that explain the relationship between variables in itemsets (LHS ∪ RHS) where the LHS => RHS (presence of the LHS leads to the presence of the RHS). The support and confidence metrics further help explain the strengths of the relationship.

Support

Support is the frequency at which a specific itemset appears in the collection of transactions, usually described as a percentage. It therefore explains the prevalence of the itemset in the transactions. While higher support means that there is more evidence that the items in the itemset appear together, low support means that the itemset appears less frequently amongst the transactions, and there is less evidence that the items from the itemset occur together.

For example, while the three-item item {bread, milk, juice} has a support of 0.2, it means that these three items occur together in the dataset 20% of the time. The three-item item {bread, milk, eggs} has a support of 0.3 indicating that these items are found together in the dataset 30% of the time. There is therefore more evidence that the items bread, milk, and eggs occur together in the dataset, further supporting the notion that these items have a stronger relationship.

Confidence

Confidence measures the strength of the relationship by determining the probability that the RHS of the association rule occurs when the LHS are present. Association rules with a higher confidence are more likely to occur while association rules with a lower confidence are less likely to occur. An association rule with a higher confidence is stronger because we can be more confident that the association rule LHS => RHS is likely to be found to be true. Confidence, or the strength of an association rule, is calculated by dividing the support of the itemset (LHS ∪ RHS) by the support of the RHS.

Confidence = (support (LHS U RHS) / (support(LHS))

Itemsets where the support of the itemset (LHS U RHS) is higher in proportion to the support of the LHS, will have a higher confidence and a higher probability of occurring. In other words, association rules where all items in the itemset have a higher frequency of occurring together in proportion to the frequency of the items in the LHS occurring together in the transaction will result in a stronger association rule.

For example, the association rule “milk, eggs => bread” is derived from the three-item itemset {milk, eggs, bread} which has a support of 0.3. The items “milk, eggs” on the other hand have a support of 0.4 since they occur 40% of the time together in the transactions. The confidence of the association rule is therefore calculated as 0.75 (0.3/0.4). The strength of this association rule is therefore quite high and we can be confident that 75% of the time, the purchase of bread will occur in the presence of milk and eggs. The strength of this association makes sense because milk and eggs are stable items as well as bread. It is therefore rational that is a customer chooses to purchase two common stable items such as milk and eggs, they are quite likely to also purchase a staple item such as bread.

The association rule “bread => milk, juice” on the other hand is calculated as 0.2 / 0.7 = 0.29. The confidence level is much lower for this association rule because the proportion of the support value for the itemset (LHS U RHS) is 0.2 and is much lower in proportion to the support value of the items in the LHS which is 0.7. The strength of the overall association rule is therefore much lower here since we can only be confident that milk and juice will be purchased 29% of the time that bread is purchased. Logically, the association rule might not be as strong because when a customer chooses to purchase bread, they may not automatically decide to purchase two different beverage.