# Association Rule Mining

10

Learning Objectives

* Understand Association Rule Mining and its business applications
* Learn how association rules are represented
* Know the Apriori algorithm and its pseudocode
* Learn how the association rule technique works in practice

### INTRODUCTION

Associate Rule Mining is a popular, unsupervised learning technique, used in businesses to help identify shopping patterns. It is also known as Market Basket Analysis. It helps find interesting relationships (affinities) between variables (items or events). Thus, it can help cross-sell related items and increase the size of a sale.

All data used in this technique is of categorical type. There is no dependent variable. It uses machine learning algorithms. The fascinating relationship between ‘sales of diapers and beers’ is how it is often explained in popular literature. This technique accepts the raw point-of-sale transaction data as input. The output produced is the description of the most frequent affinities among the items. An example of association rule would be, “a customer who bought flight tickets and hotel reservation also bought a rental car plan 60 percent of the time.”

#### Caselet: Netflix – Data Mining in Entertainment

*Netflix suggestions and recommendation engines are powered by a suite of algorithms using data of millions of customer ratings about thousands of movies. Most of these algorithms are based on the premise that similar viewing patterns represent similar user tastes. This suite of algorithms, called CineMatch, instructs Netflix’s servers to process information from its databases to determine what movies a customer is likely to enjoy. The algorithm takes into account many factors about the films themselves, the customers’ ratings, and the combined ratings of all Netflix users. The company estimates that a whopping 75 percent of viewer activity is driven by recommendations. According to Netflix, these predictions were valid around 75 percent of the time and half of Netflix users who rented CineMatch, recommended movies and gave them a five-star rating.*

*To make matches, a computer*

1. *Searches the CineMatch database for people who have rated the same movie—for example, “The Return of the Jedi”.*
2. *Determines which of those people have also rated a second movie, such as “The Matrix”.*
3. *Calculates the statistical likelihood that people who liked “Return of the Jedi” will also like “The Matrix”.*
4. *Continues this process to establish a pattern of correlations between subscribers’ ratings of many different films.*

*Netflix launched a contest in 2006 to find an algorithm that could beat CineMatch. The contest, called the Netflix Prize, promised $1 million to the first person or team to meet the accuracy goals for recommending movies based on users’ personal preferences. Each of these algorithm submissions was required to demonstrate a 10 percent improvement over CineMatch. Three years later, the $1 million prize was awarded to a team of seven people. (Source: *http://electronics.howstuffworks.com*)*

1. *Are Netflix customers being manipulated into seeing what Netflix wants them to see?*
2. *Compare this story with Amazon’s personalization engine.*

### BUSINESS APPLICATIONS OF ASSOCIATION RULES

In business environments, a pattern or knowledge can be used for many purposes. In sales and marketing, it is used for cross-marketing and cross-selling, catalog design, e-commerce site design, online advertising optimization, product pricing, and sales/promotion configurations. This analysis suggests not to put one item on sale at a time, and instead to create a bundle of products promoted as a package to sell other nonselling items.

In retail environments, it can be used for store design. Strongly associated items can be kept close together for customer convenience. Or they could be placed far from each other so that the customer has to walk the aisles and by doing so is potentially exposed to other items.

In medicine, this technique can be used for relationships between symptoms and illnesses; diagnosis and patient characteristics/treatments; genes and their functions, etc.

### REPRESENTING ASSOCIATION RULES

A generic association rule is represented between a set *X* and *Y*: *X* if *Y* [*S*%, *C*%] *X, Y* Products and/or services

*X* Left hand side (LHS)

*Y* Right hand side (RHS)

*S* Support – how often *X* and *Y* go together in the dataset, i.e., *P*(*X*  *Y*) *C* Confidence – how often Y is found, given *X*, i.e., *P*(*Y* | *X*) Example{Hotel booking, Flight booking} if {Rental Car} [30%, 60%]

[Note *P*(*X*) is the mathematical representation of the probability or chance of *X*

occurring in the dataset]

*Computation Example*

Suppose there are 1000 transactions in a dataset. There are 300 occurrences of

*X* and 150 occurrences of (*X*, *Y*) in the dataset.

Support *S* for *X* if *Y* will be *P*(*X*  *Y*) = 150/1000 = 15%

Confidence for *X* if *Y* will be *P*(*Y* | *X*) or *P*(*X*  *Y*)/*P*(*X*) = 150/300 = 50%

### ALGORITHMS FOR ASSOCIATION RULE

Not all association rules are interesting and useful, except those that are strong and occur frequently. In association rule mining, the goal is to find all the rules that satisfy the user-specified *minimum support* and *minimum confidence*. The resulting sets of rules are all the same irrespective of the algorithm used, that is, given a transaction dataset *T*, a minimum support and a minimum confidence, the set of association rules existing in *T* is *uniquely determined*.

Fortunately, there are many algorithms that are available for generating association rules. The most popular algorithms are Apriori, Eclat, FP-Growth, along with various derivatives and hybrids of the three. All the algorithms help identify the frequent itemsets, which are then converted to association rules.

### APRIORI ALGORITHM

This is the most popular algorithm used for association rule mining. The objective is to find subsets that are common to at least a minimum number of the itemsets. A frequent itemset is the one whose support is greater than or equal to minimum support threshold. The Apriori property is a downward closure property, which means that any subset of a frequent itemset is also a frequent itemset. Thus, if (A, B, C, D) is a frequent itemset, then any subset such as (A, B, C) or (B, D) is also a frequent itemset.

It uses a bottom-up approach and the size of frequent subsets is gradually in- creased, from 1-item subsets to 2-item subsets, then 3-item subsets, and so on. Groups of candidates at each level are tested against the data for minimum support.

### ASSOCIATION RULES EXERCISE

Dataset 10.1 shows a dozen sales transactions. There are six products being sold—Milk, Bread, Butter, Eggs, Cookies, and Ketchup. Transaction#1 sold Milk, Eggs, Bread and Butter. Transaction#2 sold Milk, Butter, Eggs and Ketchup and so on. The objective is to use this transaction data to find affinities between products, i.e., which products sell together often.

The support level will be set at 33 percent and the confidence level will be set at 50 percent. That means that we have decided to consider rules from only those itemsets that occur at least 33 percent of the time in the total set of transactions. Confidence level means that within those itemsets, the rules of the form *X ->* *Y* should be such that there is at least 50 percent chance of *Y* occurring based on Xoccurring.

Dataset 10.1

Transaction List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup |  |
| 4 | Milk | Bread | Butter |  |
| 5 | Bread | Butter | Cookies |  |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies |  |  |
| 8 | Milk | Bread | Butter |  |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread |  |
| 11 | Milk | Bread | Butter |  |
| 12 | Milk | Bread | Cookies | Ketchup |

First step is to compute 1-item itemsets, i.e., how often does any product sells individually.

|  |  |
| --- | --- |
| 1-itemSets | Frequency |
| Milk | 9 |
| Bread | 10 |
| Butter | 10 |
| Egg | 3 |
| Ketchup | 3 |
| Cookies | 5 |

Thus, Milk sells in 9 out of 12 transactions, Bread sells in 10 out of 12 transactions, and so on.

At every point, there is an opportunity to select itemsets of interest, and thus further analysis. Other itemsets that occur infrequently may be removed. If itemsets that occur 4 or more times out of 12 are selected, that corresponds to meeting a minimum support level of 33 percent (4 out of 12). Only 4 items make the cut. The frequent items that meet the support level of 33 percent are

|  |  |
| --- | --- |
| Frequent 1-item Sets | Frequency |
| Milk | 9 |
| Bread | 10 |
| Butter | 10 |
| Cookies | 5 |

The next step is to go for the next level of itemsets using items selected earlier, i.e., 2-item itemsets.

|  |  |
| --- | --- |
| 2-item Sets | Frequency |
| Milk, Bread | 7 |
| Milk, Butter | 7 |
| Milk, Cookies | 3 |
| Bread, Butter | 9 |
| Butter, Cookies | 3 |
| Bread, Cookies | 4 |

Thus, the sale of (Milk, Bread) is 7 times out of 12, (Milk, Butter) is 7 times, (Bread, Butter) is 9 times, and (Bread, Cookies) is 4 times.

However, only four of these transactions meet the minimum support level of 33 percent.

|  |  |
| --- | --- |
| 2-item Sets | Frequency |
| Milk, Bread | 7 |
| Milk, Butter | 7 |
| Bread, Butter | 9 |
| Bread, Cookies | 4 |

The next step is to list the next higher level of itemsets, i.e., 3-item itemsets.

|  |  |
| --- | --- |
| 3-item Sets | Frequency |
| Milk, Bread, Butter | 6 |
| Milk, Bread, Cookies | 1 |
| Bread, Butter, Cookies | 3 |

Thus the sale of (Milk, Bread, Butter) is 6 times out of 12 and (Bread, Butter, Cookies) is 3 times out of 12. One 3-item itemset meets the minimum support requirements.

3-item Sets Frequency

Milk, Bread, Butter 6

There is no room to create a 4-item itemset for this support level.

### CREATING ASSOCIATION RULES

The most interesting and complex rules at higher size itemsets start top-down with the most frequent itemsets of higher size-numbers. Association rules are created that meet the support level (>33 percent) and confidence levels (> 50 percent).

The highest level itemset that meets the support requirements is the 3-item itemset. The following itemset has a support level of 50 percent(6 out of 12).

Milk, Bread, Butter 6

This itemset could lead to multiple candidates association rules.

Start with the following rule

(Bread, Butter) -> Milk

There are a total of 12 transactions.

*X* (in this case Bread, Butter) occurs 9 times; *X*, *Y* (in this case Bread, Butter, Milk) occurs 6 times.

The support level for this rule is 6/12 = 50 percent. The confidence level for this rule is 6/9 = 67 percent. This rule meets our thresholds for support (>33 percent) and confidence (>50 percent).

Thus, the first valid association rule from this data is (Bread, Butter) -> Milk

{*S* = 50%, *C* = 67%}.

In exactly the same way, other rules can be considered for their validity.

Consider the rule (Milk, Bread) -> Butter. Out of total 12 transactions, (Milk, Bread) occurs 7 times and (Milk, Bread, Butter) occurs 6 times.

The support level for this rule is 6/12 = 50 percent. The confidence level for this rule is 6/7 = 86 percent. This rule meets our thresholds for support (>33 percent) and confidence (>50 percent).

Thus, the second valid association rule from this data is (Milk, Bread) -> Butter {*S* = 50%, *C* = 67%}.

Consider the rule (Milk, Butter) -> Bread. Out of total 12 transactions, (Milk, Butter) occurs 7 times, while (Milk, Butter, Bread) occur 6 times.

The support level for this rule is 6/12 = 50 percent. The confidence level for this rule is 6/7 = 86 percent. This rule meets our thresholds for support (>33 percent) and confidence (>50 percent).

Thus, the next valid association rule is Milk, Butter -> Bread{*S* = 50%, *C* = 86%}.

Thus, there were only three possible rules at the 3-item itemset level and all were found to be valid.

One can get to the next lower level and generate association rules at the 2-item itemset level.

Consider the following rule

Milk -> Bread;out of total 12 transactions, Milk occurs 9 times while (Milk, Bread) occurs 7 times.

The support level for this rule is 7/12 = 58 percent. The confidence level for this rule is 7/9 = 78 percent. This rule meets our thresholds for support (>33 percent) and confidence (>50 percent).

Thus, the next valid association rule is Milk -> Bread{58%, 78%}. Many such rules could be derived if needed.

Not all such association rules are interesting. The client may be interested in only the top few rules that they want to implement. The number of association rules depends upon business needs. Implementing every rule in business will require some cost and effort, with some potential of gains. The strongest of rules, with the higher support and confidence rates, should be used first, and the others should be progressively implemented later.

## Conclusion

Association rules help discover affinities between products in transactions. It helps make cross-selling recommendations much more targeted and effective. Apriori technique is the most popular technique and it is a machine learning technique.

## Questions

1. What are association rules? How do they help?
2. How many association rules should be used?
3. What are frequent itemsets?
4. How does the Apriori algorithm work?

## True/False

1. Also known as Market Basket Analysis, Association Rule Mining is a ma- chine learning algorithm.
2. Association rule mining is an unsupervised learning technique that helps find frequent patterns.
3. Not all association rules are interesting. The client may be interested in implementing only the top few rules.
4. All the data used in association rules technique is of ratio (numeric) type.
5. Given a transaction dataset *T*, a minimum support and a minimum confidence, the goal is to determine a set of rules that meet those support and confidence conditions.
6. The set of association rules existing in *T* depends upon the algorithm used.
7. A generic association rule is represented between sets *X* and *Y* as *X* -> *Y*

[*S*%, *C*%].

1. A frequent itemset can contain any number of items.
2. Apriori is the name of the most popular association rule mining technique.
3. A suite of algorithms called CineMatch helps Netflix determine which movies a customer is likely to enjoy next.