Data Warehouse & Data Mining
Unit 5

Syllabus

- Market Basket Analysis: A Motivating Example
- Frequent Item sets, Closed Item sets, and Association Rules
- Frequent Itemset Mining Methods
- Apriori Algorithm: Finding Frequent Item sets by Confined, Candidate Generation
- Generating Association Rules from Frequent Item sets

Market Basket Analysis

- **Market basket analysis (also known as association analysis) is a statistical technique used to identify relationships between products that are frequently purchased together by customers.**
- It involves analyzing customer purchase patterns and identifying items that are frequently bought together.
- The goal of market basket analysis is to discover patterns in customer behavior that can be used to improve sales and marketing strategies.
- It can be used to answer questions such as "What products are frequently purchased together?" and "Which products are often purchased by customers who buy a particular item?"
- The analysis is typically done by examining transactional data, such as sales receipts or customer order history, and using algorithms to identify relationships between products.
- The results can be used to create product recommendations, improve inventory management, and optimize store layouts and promotional campaigns.

Understanding Market Basket Analysis by Considering a Retail Use-Case

Retail – each customer purchases different set of products, different quantities, different times

MBA uses this information to:

- Identify who customers are (not by name)
- Understand why they make certain purchases
- Gain insight about its merchandise (products):
- Fast and slow movers

- Products which are purchased together
- Products which might benefit from promotion

Take action:

- Store layouts
- Which products to put on specials, promote, coupons…
- Combining all of this with a customer loyalty card it becomes even more valuable

Other Insights which can be analyzed:-

- It is also about what customers do not purchase, and why.
- If customers purchase baking powder, but no flour, what are they baking?
- If customers purchase a mobile phone, but no case, are you missing an opportunity?
- It is also about key drivers of purchases; for example, the gourmet mustard that seems to lie on a shelf collecting dust until a customer buys that particular brand of special gourmet mustard in a shopping excursion that includes hundreds of dollars' worth of other products. Would eliminating the mustard (to replace it with a better-selling item) threaten the entire customer relationship?

Some other Use-Cases on which we can perform MBA to get certain Insights.

- Items purchased on a credit card, such as rental cars and hotel rooms, provide insight into the next product that customers are likely to purchase,
- Optional services purchased by telecommunications customers (call waiting, call forwarding, DSL, speed call, and so on) help determine how to bundle these services together to maximize revenue.
- Banking products used by retail customers (money market accounts, certificate of deposit, investment services, car loans, and so on) identify customers likely to want other products.
- Unusual combinations of insurance claims can be a sign of fraud and can spark further investigation.
- Medical patient histories can give indications of likely complications based on certain combinations of treatments.

Frequent Item Sets

- Frequent item sets refer to sets of items that frequently appear together in transactions.

- **Frequent item sets refer to sets of items that appear together in a significant number of transactions or baskets. These item sets are used to identify patterns in customer purchasing behavior and are a fundamental concept in association rule mining.**

- For example, if a frequent item set consists of items A, B, and C, this means that these three items are often purchased together. This information can be used to make marketing decisions, such as placing these items in close proximity to each other in a store or creating a promotion where customers receive a discount if they purchase all three items together.

- **Real Time Example:** If a retailer discovers that customers frequently purchase both bread and milk together, they may decide to place these items near each other in the store to encourage customers to purchase both items at the same time. Alternatively, they may create a promotional offer where customers receive a discount on bread if they purchase milk at the same time.

Closed Item Sets

- In market basket analysis, a closed item set is a set of items that frequently appear together in customer transactions and is not a subset of any other item set with the same frequency. This means that it represents a complete set of items that are often purchased together and is not a part of any larger set with the same frequency.

- For example, suppose that a retailer has transactional data that shows that customers frequently purchase items A, B, C, and D together, with a support level of 20%. In addition, the data shows that items A, B, and C are often purchased together, with a support level of 25%. In this case, the closed item set would be {A, B, C}, as it has the same support level as the larger item set containing all four items but is not a subset of any other item set with the same support.

- Real Life Example: Imagine that a customer frequently purchases a burger, fries, and a drink together. This would be considered a closed item set because it represents a complete set of items that are often purchased together and is not part of any other larger set of items that are equally as popular.

- Closed item sets are useful in market basket analysis because they can help to simplify the results of association rule mining. When a dataset contains many frequent item sets, it can be challenging to identify the most significant relationships between items. Closed item sets provide a concise representation of the most important relationships between items in the dataset.

Association Rules

- **Association rules are a data mining technique used to find relationships between items in a large dataset.**
- They are often used in market basket analysis to identify items that are frequently purchased together.
- Association rules are based on the idea of finding frequent itemsets, which are sets of items that frequently occur together in the dataset.
- Once these frequent itemsets are identified, association rules can be generated to show the relationships between the items in the itemset.
- The basic idea behind association rules is to identify items that tend to be purchased together.
- For example, if a grocery store finds that customers who buy cereal are also likely to buy milk, they could use this information to create a promotional offer for customers who purchase both items together.
- The strength of an association rule is typically **measured by two metrics: support and confidence.**
- **Support** is the percentage of transactions that contain both items in the itemset, while **Confidence** is the percentage of transactions containing the first item that also contain the second item.
- Association rules can be used in a variety of applications, including marketing, sales, and recommendation systems.

Types of Association Rules

There are two main types of association rules:

1. **Positive association rules:** These are rules that indicate a positive correlation between the presence of two items in a transaction. For example, if customers who buy peanut butter are also likely to buy jelly, this would be a positive association rule.
2. **Negative association rules:** These are rules that indicate a negative correlation between the presence of two items in a transaction. For example, if customers

who buy diet soda are less likely to buy regular soda, this would be a negative association rule.

There are several algorithms used in market basket analysis to generate association rules from large datasets. Some of the most commonly used algorithms include:

1. **Apriori algorithm: This is a popular algorithm used to find frequent itemsets in a dataset. It works by generating candidate itemsets and pruning them based on their support values.**
2. FP-Growth algorithm: This is another popular algorithm used to find frequent itemsets. It works by building a tree structure that represents the dataset and uses this structure to efficiently generate frequent itemsets.
3. Eclat algorithm: This is an algorithm that works by using a vertical database representation of the dataset. It is particularly effective for sparse datasets where the number of items is much larger than the number of transactions.
4. Association rules by clustering: This is an algorithm that uses clustering techniques to identify groups of similar items in the dataset. Association rules can then be generated based on the items in each cluster.
5. CART algorithm: This is a decision tree-based algorithm that can be used to generate association rules. It works by dividing the dataset into subsets based on the values of different attributes, and then generating rules based on the relationships between the subsets.
6. Bayesian network algorithm: This is a probabilistic algorithm that uses Bayesian networks to model the relationships between items in the dataset. It can be used to generate association rules based on the probability of certain items occurring together.

Apriori Algorithm

- The Apriori algorithm is a classic algorithm used in data mining and machine learning for association rule mining.
- It is used to discover frequent itemsets from a given transaction database.
- The algorithm is based on the concept of support and confidence measures, which are used to determine the frequency of occurrence of itemsets and the strength of their association, respectively.

Working of Apriori Algorithm

Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate generation

- The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support.
- **To do this, the algorithm first looks at each individual item in the dataset and counts how many times it appears.**
- **Then, it generates a list of pairs of items, and counts how often those pairs appear together. It continues this process, generating lists of larger and larger itemsets, until it has identified all itemsets that meet a certain minimum frequency threshold.**
- **Once the algorithm has identified these frequent itemsets, we can use them to make predictions about future purchases.**
- **For example, if we know that customers who buy bread and milk often also buy eggs, we can suggest eggs to customers who have already purchased bread and milk.**
- **Overall, the Apriori algorithm is a useful tool for discovering patterns in large datasets and making predictions based on those patterns.**

Ex. Consider the following transactional dataset having min_sup = 2.

T_{ID}	Items in transaction
T1	I_1, I_2, I_5
T2	I_2, I_4
T3	I_2, I_3
T4	I_1, I_2, I_4
T5	I_1, I_3

T6	I_2, I_3
T7	I_1, I_3
T8	I_1, I_2, I_3, I_5
T9	I_1, I_2, I_3

Step 1: Scan D to find the count of each 1- itemset. It is C_1, the candidate set of 1-itemset.

Itemset	Count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

Since each itemset has its count more than or equal to min_sup, so it is L_1, the frequent 1- itemset.

Step 2: Perform Self Join over L_1 with L_1 to get C_2, the candidate set of 2-itemset.

Itemset	Count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1

$\{I_4, I_5\}$	0

Determine L_2, the frequent 2- itemset by scanning C_2.

Itemset	Count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Since the non-frequent itemset cannot be part of a frequent itemset, so we can remove them at this stage.

Step 3: Likewise determine C_3, the candidate set of 3-itemset by performing Self Join over L_2.

Reference Questions

1. What is Market Basket Analysis explain with Example
2. Explain following terms i) Frequent Item sets ii) Closed Item sets and iii) Association Rules
3. Explain Association Rule mining with example
4. Explain Apriori Algorithm with example
5. Explain the following concepts --Finding Frequent Item sets by Confined, Candidate Generation, Generating Association Rules from Frequent Item sets using Apriori Algorithm for given below example