Interview Questions

1. What is lift and why is it important in Association rules?

Ans.

In association rule mining, lift quantifies the significance of a relationship between items in transactional data. It measures the likelihood of one item being purchased given the purchase of another item, compared to the probability of this occurrence by random chance. A lift value above 1 indicates that the items are positively correlated, suggesting a meaningful association. Higher lift values imply stronger associations, which can inform decisions such as product placement or targeted marketing strategies. Understanding lift helps businesses identify valuable patterns in customer behavior, facilitating more effective decision-making and improving overall performance in sales and marketing efforts.

1. What is support and Confidence. How do you calculate them?

Ans.

1. **Support**: Support measures the frequency with which an itemset (combination of items) appears in a dataset. It indicates how popular or frequent an itemset is among all the transactions in the dataset.

Calculation: Support for an itemset is calculated by dividing the number of transactions containing that itemset by the total number of transactions in the dataset.

Example: If a dataset contains 100 transactions and 30 of them contain both items A and B, then the support for the itemset {A, B} would be 30/100 = 0.3.

1. **Confidence:** Confidence measures the reliability or certainty of a rule. It indicates the likelihood that an item B is purchased when item A is purchased. In other words, it measures the conditional probability of the consequent given the antecedent.

Calculation: Confidence for a rule (A -> B) is calculated by dividing the support for the itemset containing both A and B by the support for the itemset containing only A.

Example: If the support for {A, B} is 0.3 and the support for {A} is 0.5, then the confidence of the rule {A -> B} would be 0.3 / 0.5 = 0.6.

1. What are some limitations or challenges of Association rules mining?

Ans

1. **Spurious Correlations**: Association rules can sometimes identify spurious or coincidental correlations that are not meaningful. These rules may arise due to the nature of the data or the specificities of the algorithm, leading to potentially misleading insights.
2. **High Dimensionality**: As the number of items or attributes in the dataset increases, the number of possible itemsets and rules also grows exponentially. This can lead to a combinatorial explosion, making it computationally expensive to find all meaningful associations, especially in large datasets.
3. **Threshold Selection**: Determining appropriate thresholds for support and confidence can be challenging. Setting thresholds too low may result in many trivial or uninteresting rules, while setting them too high may cause meaningful associations to be overlooked.
4. **Sparse Data**: In real-world datasets, many itemsets may have low support, resulting in sparse data. Sparse datasets can make it difficult to find meaningful associations, as there may not be enough instances of certain itemsets to generate reliable rules.
5. **Handling Continuous Data**: Association rule mining traditionally works with categorical data, but real-world datasets often contain continuous or numerical attributes. Converting continuous data into discrete categories can lead to information loss and may affect the quality of the mined rules.
6. **Interpretability**: While association rules provide insights into relationships between items, the sheer number of rules generated can make it challenging to interpret and extract actionable knowledge from them. Additionally, some rules may be complex and difficult to understand, limiting their practical utility.
7. **Temporal Dynamics**: Association rules do not inherently capture temporal dynamics or changes in patterns over time. Analyzing temporal trends and evolving relationships between items requires additional techniques beyond traditional association rule mining.