Behind Support and Confidence

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns.

Support

In data mining, support refers to the relative frequency of an item set in a dataset. For example, if an itemset occurs in 5% of the transactions in a dataset, it has a support of 5%. Support is often used as a threshold for identifying frequent item sets in a dataset, which can be used to generate association rules. For example, if we set the support threshold to 5%, then any itemset that occurs in more than 5% of the transactions in the dataset will be considered a frequent itemset.

The support of an itemset is the number of transactions in which the itemset appears, divided by the total number of transactions. For example, suppose we have a dataset of 1000 transactions, and the itemset {milk, bread} appears in 100 of those transactions. The support of the itemset {milk, bread} would be calculated as follows:

```
Support({milk, bread}) = Number of transactions containing
                         {milk, bread} / Total number of
transactions
                       = 100 / 1000
                       = 10%
```

So the support of the itemset {milk, bread} is 10%. This means that in 10% of the transactions, the items milk and bread were both purchased.

In general, the support of an itemset can be calculated using the following formula:

Support(X) = (Number of transactions containing X) / (Total number of transactions)

where X is the itemset for which you are calculating the support.

Confidence

In data mining, confidence is a measure of the reliability or support for a given association rule. It is defined as the proportion of cases in which the association rule holds true, or in other words, the percentage of times that the items in the antecedent (the “if” part of the rule) appear in the same transaction as the items in the consequent (the “then” part of the rule).

Confidence is a measure of the likelihood that an itemset will appear if another itemset appears. For example, suppose we have a dataset of 1000 transactions, and the itemset {milk, bread} appears in 100 of those transactions. The itemset {milk}

appears in 200 of those transactions. The confidence of the rule “If a customer buys milk, they will also buy bread” would be calculated as follows:

```
Confidence("If a customer buys milk, they will also buy bread")
  = Number of transactions containing
   {milk, bread} / Number of transactions containing {milk}
  = 100 / 200
  = 50%
```

So the confidence of the rule “If a customer buys milk, they will also buy bread” is 50%. This means that in 50% of the transactions where milk was purchased, bread was also purchased.

In general, the confidence of a rule can be calculated using the following formula:

Confidence(X => Y) = (Number of transactions containing X and Y) / (Number of transactions containing X)

where X and Y are the itemsets for which you are calculating the confidence of the rule X => Y (meaning “If X, then Y”).

Support and confidence are two measures that are used in association rule mining to evaluate the strength of a rule.

Both support and confidence are used to identify strong association rules. A rule with high support is more likely to be of interest because it occurs frequently in the dataset. A rule with high confidence is more likely to be valid because it has a high likelihood of being true.

Support and Confidence

Support	Confidence
Support is a measure of the number of times an item set appears in a dataset.	Confidence is a measure of the likelihood that an itemset will appear if another itemset appears.
Support is calculated by dividing the number of transactions containing an item set by the total number of transactions.	Confidence is calculated by dividing the number of transactions containing both itemsets by the number of transactions containing the first itemset.
Support is used to identify itemsets that occur frequently in the dataset.	Confidence is used to evaluate the strength of a rule.

Support	Confidence
Support is often used with a threshold to identify itemsets that occur frequently enough to be of interest.	Confidence is often used with a threshold to identify rules that are strong enough to be of interest.
Support is interpreted as the percentage of transactions in which an item set appears.	Confidence is interpreted as the percentage of transactions in which the second itemset appears given that the first itemset appears.