Sumit suresh patil
PRN:220980728007

(i) Using the Apriori algorithm with min_sup = 60%, we can find all frequent itemsets as follows:

Find frequent 1-itemsets:
{M}, {O}, {N}, {K}, {E}, {Y}, {A}, {U}, {C}, {I}
Find frequent 2-itemsets:
{M,O}, {M,N}, {M,K}, {M,E}, {M,Y}, {O,N}, {O,K}, {O,E}, {O,Y}, {N,K}, {N,E}, {N,Y}, {K,E}, {K,Y}, {E,Y}, {M,A}, {M,K}, {M,E}, {A,K}, {A,E}, {K,E}, {M,U}, {M,C}, {M,Y}, {U,C}, {U,Y}, {C,Y}, {C,O}, {C,K}, {C,E}, {O,K}, {O,E}, {O,Y}, {K,E}, {C,I}, {O,I}, {E,I}
Find frequent 3-itemsets:
{M,O,N}, {M,O,K}, {M,O,E}, {M,O,Y}, {M,N,K}, {M,N,E}, {M,N,Y}, {M,K,E}, {M,K,Y}, {M,E,Y}, {O,N,K}, {O,N,E}, {O,N,Y}, {O,K,E}, {O,K,Y}, {O,E,Y}, {N,K,E}, {N,K,Y}, {N,E,Y}, {M,A,K}, {M,A,E}, {M,K,E}, {A,K,E}, {M,U,C}, {M,U,Y}, {M,C,Y}, {U,C,Y}, {C,O,K}, {C,O,E}, {C,K,E}, {O,K,E}, {C,O,I}, {O,E,I}, {E,I,C}
There are no frequent 4-itemsets.
There are no frequent 5-itemsets.
Therefore, the frequent itemsets with min_sup = 60% are:
{M}, {O}, {N}, {K}, {E}, {Y}, {A}, {U}, {C}, {I},
{M,O}, {M,N}, {M,K}, {M,E}, {M,Y}, {O,N}, {O,K}, {O,E}, {O,Y}, {N,K}, {N,E}, {N,Y}, {K,E}, {K,Y}, {E,Y}, {M,A}, {A,K}, {A,E}, {M,U}, {M,C}, {M,Y}, {U,C}, {U,Y}, {C,Y}, {C,O}, {C,K}, {C,E}, {O,K}, {O,E}, {O,Y}, {C,I}, {O,I}, {E,I},
{M,O,N}, {M,O,K}, {M,O,E}, {M,O,Y}, {M,N,K}, {M,N,E}, {M,N,Y}, {M,K,E}, {M,K,Y}, {M,E,Y}, {O,N,K}, {O,N,E}, {O,N,Y}, {O,K,E}, {O,K,Y}, {O,E,Y}, {N,K,E}, {N,K,Y}, {N,E,Y}, {M,A,K}, {M,A,E}, {A,K,E}, {M,U,C}, {M,U,Y}, {M,C,Y}, {U,C,Y}, {C,O,K}, {C,O,E}, {C,K,E}, {O,K,E}, {C,O,I}, {O,E,I}, {E,I,C}
(ii) List all of the strong association rules (with support s and confidence c)

To find all strong association rules with min_sup = 60% and min_conf = 80%, we can use the frequent itemsets generated in part (i). We then apply the following rules:

For each frequent itemset A, generate all non-empty subsets B of A.
For each such subset B, create the association rule B → (A - B).
Calculate the support and confidence of each rule, and keep only those rules that satisfy the minimum thresholds.
Using these steps, we obtain the following strong association rules:

{M} → {O} (s=40%, c=100%)
{O} → {M} (s=40%, c=100%)
{M} → {N} (s=40%, c=100%)
{N} → {M} (s=40%, c=100%)
{M} → {K} (s=60%, c=100%)
{K} → {M} (s=60%, c=100%)
{M} → {E} (s=60%, c=100%)
{E} → {M} (s=60%, c=100%)

{M} → {Y} (s=60%, c=100%)
{Y} → {M} (s=60%, c=100%)
{O} → {N} (s=40%, c=100%)
{N} → {O} (s=40%, c=100%)
{O} → {K} (s=40%, c=100%)
{K} → {O} (s=40%, c=100%)
{O} → {E} (s=40%, c=100%)
{E} → {O} (s=40%, c=100%)
{O} → {Y} (s=40%, c=100%)
{Y} → {O} (s=40%, c=100%)
{N} → {K} (s=40%, c=100%)
{K} → {N} (s=40%, c=100%)
{N} → {E} (s=40%, c=100%)
{E} → {N} (s=40%, c=100%)
{N} → {Y} (s=40%, c=100%)
{Y} → {N} (s=40%, c=100%)
{K} → {E} (s=60%, c=100%)
{E} → {K} (s=60%, c=100%)
{K} → {Y} (s=60%, c=100%)
{Y} → {K} (s=60%, c=100%)
{E} → {Y} (s=60%, c=100%)
{Y} → {E} (s=60%, c=100%)

Each rule is listed with its support (s) and confidence (c) values. For example, the rule {M} → {K} has a support of 60%, which means it appears in 3 out of 5 transactions, and a confidence of 100%, which means that whenever {M} appears in a transaction, {K} also appears in that transaction. Similarly, the rule {K} → {M} has a support of 60% and a confidence of 100%, which means that whenever {K} appears in a transaction, {M}

List the advantages and disadvantages of K Means clustering

- K-means clustering is a popular unsupervised machine learning algorithm used for clustering data points into groups or clusters. The algorithm works by iteratively partitioning data points into K clusters based on their similarity to each other. Here are some advantages and disadvantages of K-means clustering:

Advantages:

- K-means is a simple and easy-to-understand algorithm that is widely used in many applications.
- It is computationally efficient and works well on large datasets.
- K-means can be applied to any type of data, as long as the distance metric is well defined.
- The results of K-means are easy to interpret and can be visualized easily.

Disadvantages:

- The algorithm requires the user to specify the number of clusters (K) in advance, which may not always be known.

- K-means is sensitive to initializations, which means that different starting points can result in different cluster assignments.
- K-means assumes that clusters are spherical, equally sized, and have similar densities, which may not always be the case in real-world datasets.
- K-means can be biased towards clusters of similar sizes and may not always find the optimal solution.
- The algorithm is not suitable for datasets with high-dimensional or sparse data as the distance metric becomes less meaningful in these cases.

In summary, K-means clustering is a useful algorithm for clustering data points into groups, but it has some limitations that need to be considered when applying it to real-world datasets.

What are Type I and Type II Errors? Explain with an example.

Type I and Type II errors are two types of errors that can occur in hypothesis testing.

Type I error, also known as a false positive, occurs when the null hypothesis (H0) is rejected, but it is actually true. This means that the researcher concludes that there is a significant difference or effect when there isn't one. The probability of making a Type I error is denoted by the Greek letter alpha (α) and is set by the researcher before conducting the experiment.

Type II error, also known as a false negative, occurs when the null hypothesis is not rejected, but it is actually false. This means that the researcher fails to detect a significant difference or effect when there actually is one. The probability of making a Type II error is denoted by the Greek letter beta (β) and is related to the sample size, effect size, and the chosen level of significance.

Here's an example to illustrate these two types of errors:

Suppose a new drug is being tested to see if it lowers blood pressure. The null hypothesis is that the drug has no effect on blood pressure (H0: $\mu = 120$) and the alternative hypothesis is that it does (Ha: $\mu < 120$).

If the researcher conducts a study with a sample size of 100 and a significance level of 0.05, then the critical value is -1.645 for a one-tailed test. If the researcher finds that the sample mean is 117, which is lower than the critical value, then they reject the null hypothesis and conclude that the drug does lower blood pressure. However, there is a chance that the result is due to chance or random error, and the null hypothesis is actually true. This is a Type I error.

On the other hand, if the researcher fails to reject the null hypothesis and concludes that the drug has no effect on blood pressure, but it actually does lower blood pressure, then this is a Type II error. This could happen if the sample size is too small or if the effect size of the drug is too small to be detected with the chosen level of significance.

In both cases, the errors can have serious consequences. In the first case, patients might be given a drug that doesn't actually work, while in the second case, a drug that could have a positive effect might be discarded. Therefore, it is important to minimize the risk of both types of errors by choosing an appropriate sample size, significance level, and effect size.