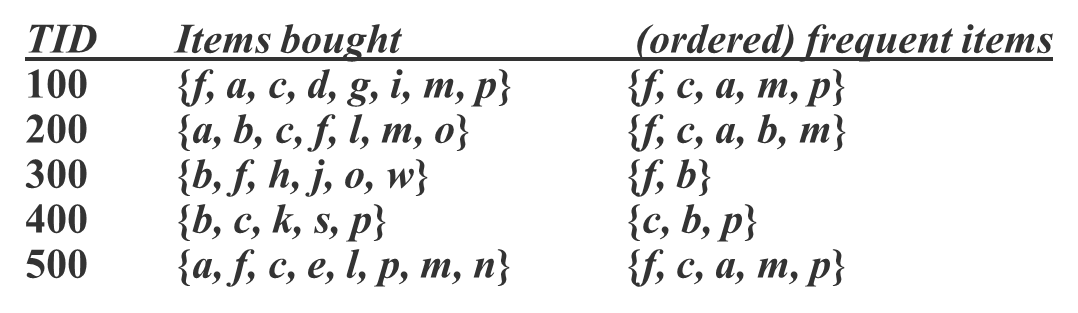
1. Construct a FP tree for the given dataset. Find the all-frequent item sets from FP tree with minimum support count 3. Draw a table for showing conditional database, conditional tree and frequent patterns at each step.
2. Using the following dataset, create two clusters (K=2) using K-Mean Clustering Method, taking K1=P2 and K2=P8 as initial centroids. The distance function between two points a*=(x1, y1)* and b*=(x2, y2)* is defined as: *ρ(a, b) = |x1 – x2| + |y1 – y2|*. Show the clusters after second iteration. [*05 Marks*]

|  |  |  |
| --- | --- | --- |
| Pt. | X | Y |
| P1 | 2 | 3 |
| P2 | 3 | 1 |
| P3 | 4 | 2 |
| P4 | 11 | 5 |
| P5 | 12 | 4 |
| P6 | 12 | 6 |
| P7 | 7 | 5 |
| P8 | 8 | 4 |
| P9 | 8 | 6 |

1. What is ‘Confusion Matrix? How it can be composed and used for model evaluation? In case of class imbalance problem the accuracy measure is assumed to be biased; what other measures could be employed to deal with such kind of problem.

Answer:

Confusion matrix is a great measure used for solving classification problems and measuring the accuracy. It can be applied to different classification problems such as

* Binary classification
* Multiclass classification

It uses a table to define the matrix such as

* False Negative
* False Positive
* True Negative
* True Positive

It represents counts from **actual** and **predicted** values.

It is a table that allows you to visualize the performance of a classification model. You can also use the information in it to calculate measures that can help you determine the usefulness of the model. Rows represent predicted classifications, while columns represent the true classes from the data. It contains four outcomes produced by a binary classifier. Various measures, such as error-rate, accuracy, specificity, sensitivity, and precision, are derived from the confusion matrix.

Accuracy can be misleading if used with imbalanced datasets, and therefore there are other metrics based on confusion matrix which can be useful for evaluating performance.

There are two groups of metrics that may be useful for imbalanced classification because they focus on one class; they are sensitivity-specificity and precision-recall.