Discretization in data mining

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy.

In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.

There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization.

Supervised discretization refers to a method in which the class data is used.

Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

| Age | 1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77 |
| --- | --- |

Table before Discretization

| Attribute | Age | Age | Age | Age |
| --- | --- | --- | --- | --- |
|  | 1,5,4,9,7 | 11,14,17,13,18,19 | 31,33,36,42,44,46 | 70,74,77,78 |
| After Discretization | Child | Young | Mature | Old |

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level.

**Some Famous techniques of data discretization Histogram analysis**

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

**Binning**

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

**Cluster Analysis**

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

**Data discretization using decision tree analysis**

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

**Data discretization using correlation analysis**

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

**Data discretization and concept hierarchy generation**

The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance. In other words, we can say that a hierarchy concept refers to a sequence of mappings with a set of more general concepts to complex concepts. It means mapping is done from low-level concepts to high-level concepts. For example, in computer science, there are different types of hierarchical systems. A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model. There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping.

Let's understand this concept hierarchy for the dimension location with the help of an example.

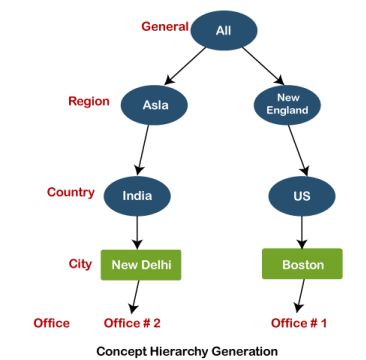
A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

**Top-down mapping**

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

**Bottom-up mapping**

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.

**Data discretization and binarization in data mining**

Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. In contrast, data binarization is used to transform the continuous and discrete attributes into binary attributes.

**Why is Discretization important?**

As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.