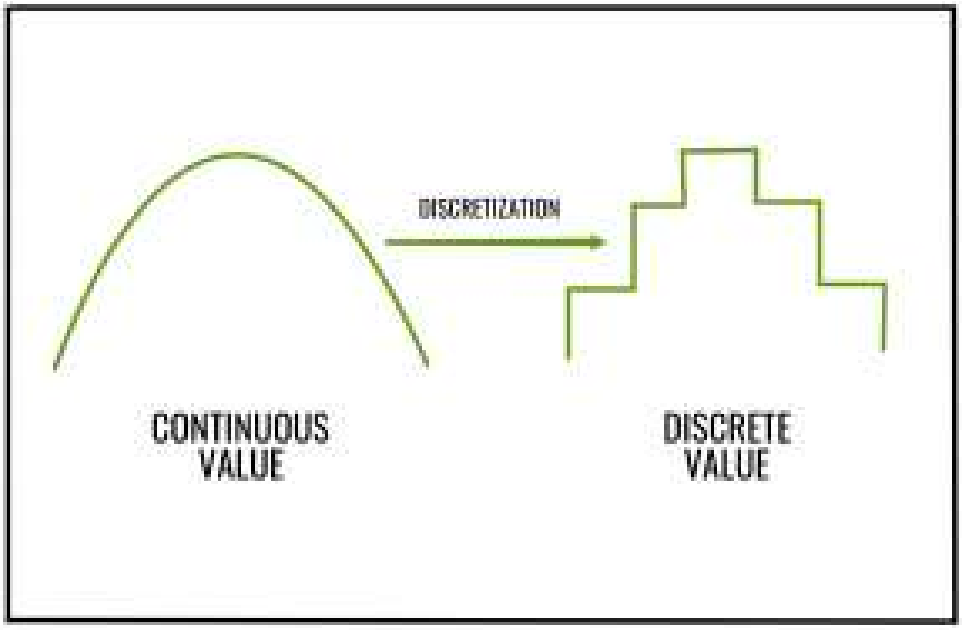
Data discretization in data mining



Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy.

Data discretization example

we have an attribute of age with the following values.

Age	10,11,13,14,17,19,30, 31, 32, 38, 40, 42,70 , 72, 73, 75

Table: Before discretization

Attribute	Age	Age	Age
	10,11,13,14,17,19,	30, 31, 32, 38, 40, 42	70 , 72, 73, 75
After Discretization	Young	Mature	Old

What are some famous techniques of data discretization?

1. Histogram analysis: **Histogram** is a plot used to present the underlying frequency distribution of a set of continuous data. The histogram helps the inspection of the data for the distribution of the data. For example normal distribution representation, outliers, and skewness representation, etc.
2. Binning: **Binning** is a data smoothing technique and its helps to group a huge number of continuous values into a smaller number of bins. For example, if we have data about a group of students, and we want to arrange their marks into a smaller number of marks intervals by making the bins of grades. One bin for grade A, one for grade B, one for C, one for D, and one for F Grade.
3. **Correlation analysis:** Cluster analysis is commonly known as clustering. Clustering is the task of grouping similar objects in one group, commonly called clusters. All different objects are placed in different clusters.
4. Clustering analysis
5. Decision tree analysis

Data discretization and concept hierarchy generation

A concept hierarchy represents a sequence of mappings with a set of more general concepts to specialized concepts. Similarly mapping from low-level concepts to higher-level concepts. In other words, we can say top-down mapping and bottom-up mapping.

Top-down mapping

Top-down mapping starts from the top with general concepts and moves to the bottom to the specialized concepts.

Bottom-up mapping

Bottom-up mapping starts from the Bottom with specialized concepts and moves to the top to the generalized concepts.

What is Binning in Data Mining?

- The original data values are divided into small intervals known as **bins,** and then they are replaced by a general value calculated for that bin.
- This has a soothing effect on the input data and may also reduce the chances of over fitting in the case of small datasets.

Why is Binning Used?

Binning or discretization is used to transform a continuous or numerical variable into a categorical feature.