Semester : V Subject :DWM Academic Year: 2023 - 2024

Module 2

The attribute can be defined as a field for storing the data that represents the characteristics of a data object. The attribute is the property of the object. The attribute represents different features of the object. For example, hair color is the attribute of a lady. Similarly, rollno, and marks are attributes of a student. An attribute vector is commonly known as a set of attributes that are used to describe a given object.
Type of attributes
We need to differentiate between different types of attributes during Data-preprocessing. So firstly, we need to differentiate between qualitative and quantitative attributes.
1. **Qualitative Attributes** such as Nominal, Ordinal, and Binary Attributes.
2. **Quantitative Attributes** such as Discrete and Continuous Attributes.
There are different types of attributes. some of these attributes are mentioned below;

Nominal Attributes – Nominal defines associating with names. The values of a nominal attribute are symbols or names of things. Each value defines some type of category, code, or state, etc. Nominal attributes are defined as categorical. The values do not have any significant order. In computer science, the values are also called an enumerations.

Binary Attributes – A binary attribute is a nominal attribute with only two elements or states such as 0 or 1, where 0 generally defines that the attribute is absent, and 1 defines that it is present. Binary attributes are defined as Boolean if the two states equivalent to true and false.

A binary attribute is symmetric if both of its states are same valuable and produce the same weight. There is no preference on which results must be coded as 0 or 1. An example can be the attribute gender having the states male and female.

A binary attribute is asymmetric if the results of the states are not similarly important, including the positive and negative results of a medical test for HIV. By convention, it can code the most essential result, which is generally the nearest one, by 1 (e.g., HIV positive) and the different by 0 (e.g., HIV negative).

Ordinal Attributes – An ordinal attribute is an attribute with possible values that have a significant order or ranking between them, but the magnitude among successive values is not known.

Numeric Attributes – A numeric attribute is quantitative. It is a measurable quantity, defined in integer or real values. It can be interval-scaled or ratio-scaled.

Ratio Data:

This type of data is similar to interval data, but with an absolute zero point. In ratio data, it is possible to compute ratios of two values, and this makes it possible to make meaningful comparisons.

Examples of ratio data include height, weight, and income. Ratio data is used in data mining for prediction and association rule mining tasks.

Data Quality: Why do we preprocess the data?

Data preprocessing is an essential step in data mining and machine learning as it helps to ensure the quality of data used for analysis. There are several factors that are used for data quality assessment, including:

1.Incompleteness:

This refers to missing data or information in the dataset. Missing data can result from various factors, such as errors during data entry or data loss during transmission. Preprocessing techniques, such as imputation, can be used to fill in missing values to ensure the completeness of the dataset.

2.Inconsistency:

This refers to conflicting or contradictory data in the dataset. Inconsistent data can result from errors in data entry, data integration, or data storage. Preprocessing techniques, such as data cleaning and data integration, can be used to detect and resolve inconsistencies in the dataset.

3.Noise:

This refers to random or irrelevant data in the dataset. Noise can result from errors during data collection or data entry. Preprocessing techniques, such as data smoothing and outlier detection, can be used to remove noise from the dataset.

4.Outliers:

Outliers are data points that are significantly different from the other data points in the dataset. Outliers can result from errors in data collection, data entry, or data transmission. Preprocessing techniques, such as outlier detection and removal, can be used to identify and remove outliers from the dataset.

5.Redundancy:

Redundancy refers to the presence of duplicate or overlapping data in the dataset. Redundant data can result from data integration or data storage. Preprocessing techniques, such as data deduplication, can be used to remove redundant data from the dataset.

5.Data format:

This refers to the structure and format of the data in the dataset. Data may be in different formats, such as text, numerical, or categorical. Preprocessing techniques, such as data transformation and normalization, can be used to convert data into a consistent format for analysis.