Data Objects and Attribute types

Data Objects

- Data sets are made up of data objects. A data object represents an entity—in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a university database, the objects may be students, professors, and courses. Data objects are typically described by attributes.
- Data objects can also be referred to as samples, examples, instances, data points, or objects. If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

Attribute

An attribute is a data field, representing a characteristic or feature of a data object. The distribution of data involving one attribute (or variable) is called univariate. A bivariate distribution involves two attributes, and so on.

- E.g., customer _ID, name, address

- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled
- **Nominal-** categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
- Size = {small, medium, large}, grades, army rankings.

Numeric Attribute Types

A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

- Interval -Scaled Attributes
    - Measured on a scale of equal-sized units
    - The values of interval-scaled attributes have order and can be positive, 0, or negative.E.g., temperature in C˚or F˚, calendar dates

- Ratio -Scaled Attributes
    - It is a numeric attribute with an Inherent zero-point.
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - E.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- Discrete Attribute
    - Has only a finite or countably infinite set of values
        - E.g., zip codes, profession, or the set of words in a collection of documents
    - Sometimes, represented as integer variables
    - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute
    - Has real numbers as attribute values
        - E.g., temperature, height, or weight
    - Practically, real values can only be measured and represented using a finite number of digits.
    - Continuous attributes are typically represented as floating-point variables.