**Data Objects and Attribute Types**

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.

**What Is an Attribute?**

An **attribute** is a data field, representing a characteristic or feature of a data object. The term *dimension* is commonly used in data warehousing. Attributes describing a customer object can include, for example, *customer ID*, *name*, and *address*. Observed values for a given attribute are known as *observations*. A setof attributes used to describe a given object is called an *attribute vector* (or *feature vector*). The distribution of data involving one attribute (or variable) is called *univariate*.A *bivariate* distribution involves two attributes, and so on.

The type of an attribute is determined by the set of possible values—nominal, binary, ordinal, or numeric—the attribute can have. In the following subsections, we introduce each type.

**Nominal Attributes**

Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nomi-nal attributes are also referred to as categorical. The values do not have any meaningful order. In computer science, the values are also known as enumerations.

Example: Suppose that hair color and marital status are two attributes describing person objects. In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white. The attribute marital status can take on the values single, married, divorced, and widowed. Both hair color and marital status are nominal attributes. Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on.

**Binary Attributes**

A **binary attribute** is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as **Boolean** if the two states correspond to *true* and *false*.

**Example 2.2 Binary attributes.** Given the attribute*smoker*describing a*patient*object, 1 indicatesthat the patient smokes, while 0 indicates that the patient does not. Similarly, suppose the patient undergoes a medical test that has two possible outcomes. The attribute *medical test* is binary, where a value of 1 means the result of the test for the patientis positive, while 0 means the result is negative.

A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1. One such example could be the attribute *gender* having the states *male* and *female*.

A binary attribute is **asymmetric** if the outcomes of the states are not equally impor-tant, such as the *positive* and *negative* outcomes of a medical test for HIV. By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., *HIV* *positive*) and the other by 0 (e.g., *HIV negative*).

**Ordinal Attributes**

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

**Example 2.3 Ordinal attributes.** Suppose that*drink size*corresponds to the size of drinks available ata fast-food restaurant. This nominal attribute has three possible values: *small, medium*, and *large*. The values have a meaningful sequence (which corresponds to increasing drink size);

**Numeric Attributes**

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**

**Interval-scaled attributes** are measured on a scale of equal-size units. The values ofinterval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the *difference* between values.

**Example 2.4 Interval-scaled attributes.** A*temperature*attribute is interval-scaled. Suppose that wehave the outdoor *temperature* value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to *temperature*. In addition, we can quantify the difference between values. For example, atemperature of 20 C is five degrees higher than a temperature of 15 C. Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.

**Ratio-Scaled Attributes**

A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.

**Example 2.5 Ratio-scaled attributes.** Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K)temperature scale has what is considered a true zero-point (0 K D 273.15 C): It is the point at which the particles that comprise matter have zero kinetic energy. Other examples of ratio-scaled attributes include *count* attributes such as *years of experience* (e.g., the objects are employees) and *number of words* (e.g., the objects are documents). Additional examples include attributes to measure weight, height, latitude and longitude coordinates.

**Discrete versus Continuous Attributes**

A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes *hair color*, *smoker*, *medical test*, and *drink size* each have a finite number of values, and so are discrete.

An attribute is *countably infinite* if the set of possible values is infinite but the values can be put in a one-to-one correspondence with natural numbers. For example, the attribute *customer ID* is countably infinite. The number of customers can grow to infinity, but in reality, the actual set of values is countable. Zip codes are another example.

If an attribute is not discrete, it is **continuous**. In practice, real values are represented using a finite number of digits. Continuous attributes are typically represented as floating-point variables.