# Data Objects and Attribute Types

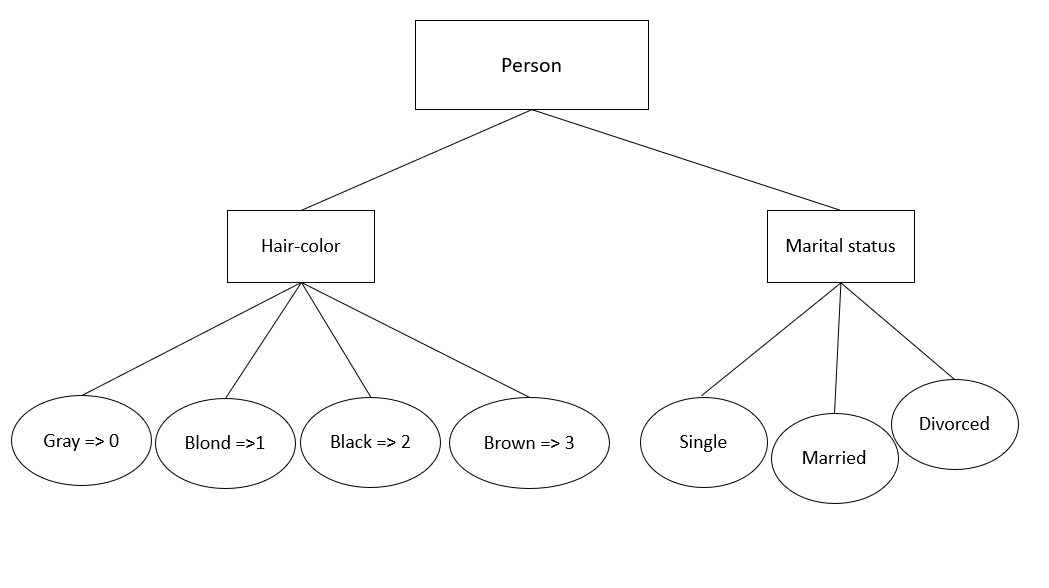
* Các tập dữ liệu được tạo thành từ các đối tượng dữ liệu.
* Một đối tượng dữ liệu biểu thị một thực thể — trong cơ sở dữ liệu bán hàng, các đối tượng có thể là khách hàng, mặt hàng trong cửa hàng và doanh số; trong cơ sở dữ liệu y tế, các đối tượng có thể là bệnh nhân; trong cơ sở dữ liệu trường đại học, các đối tượng có thể là sinh viên, giáo sư và các khóa học.
* Các đối tượng dữ liệu thường được mô tả bằng các thuộc tính. Các đối tượng dữ liệu cũng có thể được gọi là samples(mẫu), examples (ví dụ), instances(trường hợp), data points, or objects.
* Nếu các đối tượng dữ liệu được lưu trữ trong cơ sở dữ liệu, chúng là các bộ dữ liệu. Nghĩa là, các hàng của cơ sở dữ liệu tương ứng với các đối tượng dữ liệu và các cột tương ứng với các thuộc tính.

## 1. What is an Attribute?

* An attribute is a data field that represents a characteristic or feature of a data object. Different disciplines use various terms such as dimension, feature, variable, or attribute. A set of attributes used to describe a particular object is called an attribute vector.
* The basic types of attributes:
* Based on properties: Nominal, binary, ordinal, and numeric.
* Based on value domain: Discrete and continuous.
* Thuộc tính là một trường dữ liệu, đại diện cho một đặc điểm hoặc tính chất của một đối tượng dữ liệu. Trong các chuyên ngành khác nhau sẽ dùng những thuật ngữ khác nhau như chiều, đặc trưng, biến hoặc thuộc tính.Một tập hợp các thuộc tính được sử dụng để mô tả một đối tượng nhất định được gọi là vector thuộc tính.
* Những loại thuộc tính cơ bản:
* Dựa trên tính chất: Định danh, nhị phân, thứ tự, số.
* Dựa trên miền giá trị: Rời rạc và liên tục.

## 2. Nominal Attributes

* Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things
* Nominal attributes, as related to names, refer to categorical data where the values represent different categories or labels without any inherent order or ranking. These attributes are often used to represent names or labels associated with objects, entities, or concepts.
* Each value represents some kind of category, code, or state, and so nominal 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 (liệt kê)
* It is possible to represent such symbols or “names” with numbers ( Another example is customer ID, with possible values that are all numeric )
* Example:



* However, in such cases, the numbers are not intended to be used quantitatively
* Example: It makes no sense to subtract one customer ID number from another, unlike, say, subtracting an age value from another (where age is a numeric attribute). ( Sẽ không có ý nghĩa gì khi chúng ta sử dụng ID của khách hàng này để trừ đi ID của 1 khách hàng khác, nó sẽ không giống như trừ đi 1 giá trị tuổi, Mặc dù thuộc tính danh nghĩa có thể có các giá trị nguyên, nhưng nó không được coi là thuộc tính số vì các số nguyên không nhằm mục đích sử dụng định lượng )
* Dịch sang Tiếng Việt:

Thuộc tính định danh là thuộc tính "liên quan đến tên". Các giá trị của thuộc tính định danh là ký hiệu hoặc tên của sự vật. Mỗi giá trị đại diện cho một loại danh mục, mã hoặc trạng thái nào đó, và do đó các thuộc tính định danh cũng được gọi là thuộc tính phân loại. Các giá trị không phân biệt thứ tự. Trong khoa học máy tính, các giá trị cũng được gọi là liệt kê.

Mặc dù chúng tôi đã nói rằng các giá trị của một thuộc tính định danh là các ký hiệu hoặc "tên của sự vật", nhưng có thể biểu diễn các ký hiệu hoặc "tên" đó bằng số. Ví dụ, với màu tóc, chúng ta có thể gán mã 0 cho màu đen, 1 cho màu nâu, v.v. Một ví dụ khác là ID tùy chỉnh, với các giá trị có thể đều là số. Tuy nhiên, trong những trường hợp như vậy, các con số không nhằm mục đích được sử dụng để định lượng. Đó là, các phép toán trên các giá trị của các thuộc tính danh nghĩa là không có ý nghĩa. Thật vô nghĩa khi trừ một số ID khách hàng này khỏi một số ID khách hàng khác, không giống như trừ đi giá trị tuổi từ một số khác (trong đó tuổi là một thuộc tính số). Mặc dù một thuộc tính danh nghĩa có thể có số nguyên làm giá trị, nó không được coi là thuộc tính số vì các số nguyên không có nghĩa là được sử dụng định lượng.

Bởi vì các giá trị thuộc tính danh nghĩa không có bất kỳ thứ tự có ý nghĩa nào về chúng và không định lượng, nên không có ý nghĩa gì khi tìm giá trị trung bình cho một thuộc tính như vậy, cho một tập hợp các đối tượng. Tuy nhiên, một điều đáng quan tâm là giá trị xảy ra phổ biến nhất của thuộc tính. Giá trị này, được gọi là chế độ (mode) , là một trong những thước đo của xu hướng trung tâm.

## 3. Binary attributes

* Thuộc tính nhị phân là một loại thuộc tính định danh chỉ có hai danh mục hoặc trạng thái: 0 hoặc 1. Trong đó, 0 thường có nghĩa là thuộc tính vắng mặt và 1 có nghĩa là thuộc tính hiện diện. Các thuộc tính nhị phân được gọi là Boolean nếu hai trạng thái tương ứng đúng và sai.

A binary attribute is a nominal attribute with only two categories or states: 0 or 1

VD: Với thuộc tính smoker mô tả đối tượng bệnh nhân, 1 cho biết bệnh nhân hút thuốc, trong khi 0 cho biết bệnh nhân không hút thuốc. Tương tự, giả sử bệnh nhân trải qua một xét nghiệm y tế có hai kết quả có thể. Thuộc tính medical test là nhị phân, trong đó giá trị 1 có nghĩa là kết quả xét nghiệm của bệnh nhân là dương tính, trong khi 0 có nghĩa là kết quả là âm tính.

* Một thuộc tính nhị phân được gọi là đối xứng nếu cả hai trạng thái của nó đều có giá trị và trọng lượng bằng nhau; nghĩa là không có ưu tiên cho kết quả nào nên được mã hóa là 0 hoặc 1. Một ví dụ như vậy có thể là thuộc tính gender có các trạng thái male và female.

A binary attribute is symmetric if both of its states are equally valuable and carry the same weight

* Một thuộc tính nhị phân được gọi là không đối xứng nếu giá trị của các trạng thái có độ ưu tiên không bằng nhau, chẳng hạn như kết quả dương tính và âm tính của xét nghiệm HIV. Theo quy ước, chúng ta mã hóa kết quả quan trọng nhất, thường là kết quả hiếm nhất, bằng 1 (ví dụ: HIV dương tính) và kết quả còn lại bằng 0 (ví dụ: HIV âm tính).

A binary attribute is asymmetric if the outcomes of the states are not equally important

## 4. Ordinal Attributes

* Định nghĩa: Thuộc tính thứ bậc là một thuộc tính với những giá trị khả thi, có một trình tự có ý nghĩa hoặc xếp hạng giữa chúng. (An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them)
* Thuộc tính thứ bậc hữu ích cho việc ghi lại các đánh giá chủ quan của chất lượng mà không thể đo đạc một cách khách quan vì chúng thường được sử dụng trong các cuộc khảo sát đánh giá. (Ordinal attributes are useful for registering subjective assessments of qualities that cannot be measured objectively; thus ordinal attributes are often used in surveys for ratings.)
* Những thuộc tính thứ bậc thường được sử dụng trong các khảo sát cho việc đánh giá. (Ordinal attributes are often used in surveys for evaluations.)
* Thuộc tính thứ bậc có thể thu được từ việc rời rạc hóa số lượng bằng cách chia dải giá trị thành một số lượng hữu hạn các danh mục có thứ bậc. (Ordinal attributes may also be obtained from the discretization of numeric quantities by splitting the value range into a finite number of ordered categories)
* Khuynh hướng trung tâm của một thuộc tính thứ bậc có thể được biểu diễn bằng giá trị mode và median (giá trị trung bình trong một dãy có thứ tự), nhưng không thể xác định giá trị mean. (The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence, but the mean cannot be defined.)
* Thuộc tính thứ bậc là thuộc tính định tính. (Ordinal attributes are qualitative). Mô tả một đặc điểm của đối tượng mà không đưa ra kích thước hay số lượng cụ thể. Các giá trị của những thuộc tính định tính này thường là các từ đại diện cho các danh mục. Nếu sử dụng số nguyên, chúng đại diện cho mã máy tính của các danh mục, thay vì đại diện cho các số lượng có thể đo lường.
* Ví dụ:
* Kích thước đồ uống trong một nhà hàng, trong một quán ăn nhanh có thể có các giá trị như: nhỏ, vừa, lớn.
* Điểm số phân loại sinh viên của các trường đại học được chia theo từng mức điểm của mỗi cá nhân, các giá trị thứ bậc có thể là: A+, A, B+, B, …
* Trong lĩnh vực giáo dục, các cấp bậc của một con người có các giá trị như: giáo sư, phó giáo sư, tiến sĩ, thạc sĩ, kỹ sư, cử nhân, nghiên cứu sinh, …
* Các hàm, cấp bậc của các cán bộ phục vụ trong quân đội của một quốc gia cũng được định nghĩa theo thuộc tính thứ bậc: tướng, tá, úy, binh nhì, binh nhất, …

## 5. Numeric Attributes

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

· Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-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.

· 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.

Examples of interval-scaled attributes:

* Temperature: While we can compare temperatures (e.g., 20°C is 5 degrees higher than 15°C), there's no absolute zero that indicates "no temperature."
* Calendar dates: Years can be compared, and differences calculated (e.g., 2010 is 8 years after 2002), but there's no true "beginning of time" at year 0

Examples of ratio-scaled attributes:

* Temperature: in Kelvin (K): Absolute zero (0 K) represents the point at which particles have no kinetic energy.
* Count attributes: Years of experience, number of words, etc.
* Physical measurements: Weight, height, latitude, longitude.
* Monetary quantities: Dollars, euros, etc.

Comparison

· Interval Scale:

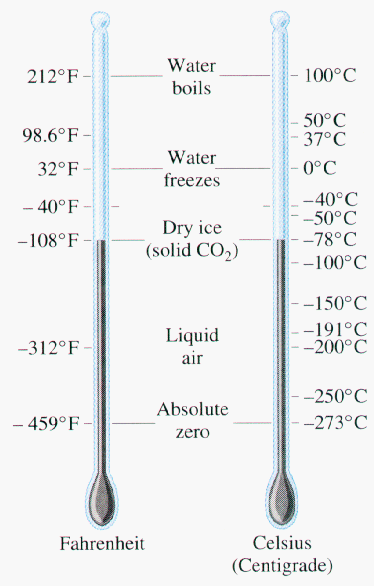
o Allows for the comparison of intervals between values.

o Does not have a true zero point.

o Enables the use of arithmetic operations like addition and subtraction.

o Commonly used in psychology, social sciences, and market research.

o Examples include temperature scales (Celsius, Fahrenheit) and Likert scales.



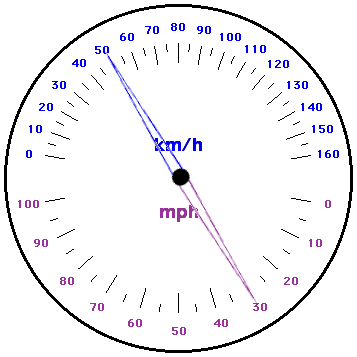
· Ratio Scale:

o Allows for the comparison of intervals between values.

o Has a true zero point.

o Enables the use of all arithmetic operations, including multiplication and division.

o Widely used in physical sciences, engineering, finance, and economics.



o Examples include weight, height, time, and precise measurements.

<https://thisvsthat.io/interval-vs-ratio>

Thuộc tính số là thuộc tính định lượng, tức là một đại lượng có thể đo được, được biểu diễn bằng giá trị nguyên hoặc thực. Thuộc tính này có thể được định lượng theo thang đo (Interval-scaled) khoảng hoặc thang đo tỷ lệ (Ratio-scaled).

· Thang đo khoảng: Các thuộc tính được đo theo thang đo khoảng được đo bằng các đơn vị có kích thước bằng nhau. Các giá trị của thuộc tính thang đo khoảng có thứ tự và có thể dương, 0 hoặc âm. Do đó, ngoài việc cung cấp thứ hạng cho các giá trị, các thuộc tính này cho phép chúng ta so sánh và định lượng sự khác biệt giữa các giá trị.

· Thang đo tỷ lệ: Thuộc tính thang đo tỷ lệ là một thuộc tính số có điểm 0 cố hữu, điểm bắt đầu (zeros point). Điều đó có nghĩa là nếu một phép đo được định lượng theo thang đo tỷ lệ, chúng ta có thể nói về một giá trị là bội số (hoặc tỷ lệ) của một giá trị khác. Ngoài ra, các giá trị được sắp xếp theo thứ tự, và chúng ta cũng có thể tính toán sự khác biệt giữa các giá trị, cũng như trung bình, trung vị và mode.

Giải thích:

· Thuộc tính số (Numeric attribute): Thuộc tính có thể được biểu diễn bằng số, ví dụ như độ tuổi, chiều cao, cân nặng.

· Định lượng (Quantitative): Có thể đo lường và biểu diễn bằng số.

· Thang đo khoảng (Interval scale): Thang đo cho phép so sánh khoảng cách giữa các giá trị nhưng không có điểm 0 tuyệt đối. Ví dụ: nhiệt độ đo theo độ Celsius.

· Thang đo tỷ lệ (Ratio scale): Thang đo có điểm 0 tuyệt đối, cho phép so sánh tỷ lệ giữa các giá trị. Ví dụ: chiều cao, cân nặng, tuổi.

So sánh:

| Thang đo khoảng  (Interval-scaled) | Thang đo tỷ lệ  (Ratio-scaled) |
| --- | --- |
| Cho phép so sánh giữa các giá trị | |
| Không có điểm bắt đầu tuyệt đối (zero point) | Có điểm bắt đầu tuyệt đối |
| Thực hiện được phép cộng và trừ | Thực hiện được phép cộng, trừ, nhân chia |
| Sử dụng trong nghiên cứu xã hội, thị trường, tâm lý | Sử dụng trong vật lý, kỹ thuật, tài chính, kinh tế |

Ví dụ:

· Thang đo khoảng:

o Nhiệt độ: Khi đo bằng độ C hoặc độ F, chúng ta có thể nói rằng 30 độ C nóng hơn 20 độ C và khoảng cách giữa hai giá trị này là 10 độ. Tuy nhiên, 0 độ C không có nghĩa là không có nhiệt độ.

o Điểm số IQ: Khoảng cách giữa IQ 100 và 110 là bằng nhau với khoảng cách giữa IQ 110 và 120. Tuy nhiên, IQ 0 không có nghĩa là không có trí thông minh.

o Thời gian: Khi đo theo giờ, phút, giây, chúng ta có thể nói rằng 3 giờ chiều muộn hơn 2 giờ chiều và khoảng cách giữa hai thời điểm này là 1 giờ. Tuy nhiên, 0 giờ không có nghĩa là không có thời gian.

· Thang đo tỷ lệ:

o Độ cao: Một người cao 180cm cao gấp đôi người cao 90cm. Điểm bắt đầu tuyệt đối là 0cm (không có chiều cao).

o Khối lượng: Một vật nặng 10kg nặng gấp đôi vật nặng 5kg. Điểm bắt đầu tuyệt đối là 0kg (không có khối lượng).

o Thu nhập: Một người có thu nhập 10 triệu đồng có thu nhập gấp đôi người có thu nhập 5 triệu đồng. Điểm bắt đầu tuyệt đối là 0 đồng (không có thu nhập).

o Độ dài: Một đoạn thẳng dài 10cm dài gấp đôi đoạn thẳng dài 5cm. Điểm bắt đầu tuyệt đối là 0cm (không có độ dài).

=> Kết luận: Hiểu được sự khác biệt giữa thang đo khoảng và thang đo tỷ lệ là rất quan trọng trong data mining, vì nó ảnh hưởng đến cách chúng ta phân tích và giải thích dữ liệu.

1. Thuộc tính Rời rạc và Liên tục

Dịch sách:

* Trong phần trình bày của chúng tôi, chúng tôi đã phân loại các thuộc tính thành danh định, nhị phân, thứ tự và số. Có nhiều cách để phân loại các thuộc tính. Các loại này không loại trừ lẫn nhau.
* Các thuật toán phân loại phát triển từ lĩnh vực học máy thường nói về các thuộc tính là rời rạc hoặc liên tục. Mỗi loại có thể được xử lý khác nhau.
* Thuộc tính rời rạc có một tập hợp giá trị hữu hạn hoặc vô hạn đếm được, có thể hoặc không thể biểu diễn dưới dạng số nguyên. Các thuộc tính như màu tóc, người hút thuốc, kết quả kiểm tra y tế và kích cỡ đồ uống đều có một số lượng giá trị hữu hạn, do đó là rời rạc. Lưu ý rằng các thuộc tính rời rạc có thể có giá trị số, chẳng hạn như 0 và 1 đối với các thuộc tính nhị phân hoặc các giá trị từ 0 đến 110 đối với thuộc tính tuổi. Một thuộc tính là vô hạn đếm được nếu tập hợp các giá trị khả dĩ là vô hạn nhưng các giá trị có thể được đặt vào một tương ứng một-một với các số tự nhiên. Ví dụ, thuộc tính mã khách hàng là vô hạn đếm được. Số lượng khách hàng có thể tăng lên vô hạn, nhưng trong thực tế, tập hợp các giá trị thực tế là có thể đếm được (khi các giá trị có thể được đặt vào một tương ứng một-một với tập hợp các số nguyên). Mã zip là một ví dụ khác.
* Nếu một thuộc tính không phải là rời rạc, thì đó là liên tục. Các thuật ngữ "thuộc tính số" và "thuộc tính liên tục" thường được sử dụng thay thế cho nhau trong tài liệu. (Điều này có thể gây nhầm lẫn vì, theo nghĩa cổ điển, các giá trị liên tục là các số thực, trong khi các giá trị số có thể là số nguyên hoặc số thực.) Trong thực tế, các giá trị thực được biểu diễn bằng một số lượng chữ số hữu hạn. Các thuộc tính liên tục thường được biểu diễn dưới dạng các biến dấu phẩy động.

Tóm tắt:

* Phân loại thuộc tính: Có nhiều cách phân loại thuộc tính, nhưng hai cách phổ biến là:
  + Dựa trên tính chất: Định danh, nhị phân, thứ tự, số.
  + Dựa trên miền giá t rị: Rời rạc và liên tục.
* Thuộc tính rời rạc:
  + Có tập giá trị hữu hạn hoặc vô hạn đếm được.
  + Có thể biểu diễn dưới dạng số nguyên hoặc không.
  + Ví dụ: màu sắc, giới tính, mã khách hàng.
* Thuộc tính liên tục:
  + Có thể có bất kỳ giá trị nào trong một khoảng số thực.
  + Thường được biểu diễn dưới dạng số dấu phẩy động.
  + Ví dụ: chiều cao, cân nặng, nhiệt độ.
* Attribute Classification: There are various ways to classify attributes, but two common approaches are:
  + Based on properties: Nominal, binary, ordinal, and numeric.
  + Based on value domain: Discrete and continuous.
* Discrete Attributes:
  + Have a finite or countably infinite set of values.
  + Can be represented as integers or not.
  + Examples: color, gender, customer ID.
* Continuous Attributes:
  + Can take on any value within a real number range.
  + Often represented as floating-point numbers.
  + Examples: height, weight, temperature.
* Quan hệ giữa các loại thuộc tính:
  + Các loại thuộc tính không loại trừ lẫn nhau.
  + "Thuộc tính số" và "thuộc tính liên tục" thường được sử dụng thay thế cho nhau.
* Quan trọng:
  + Cách phân loại thuộc tính ảnh hưởng đến việc lựa chọn thuật toán phân loại.
  + Hiểu rõ tính chất của từng loại thuộc tính giúp chúng ta xử lý dữ liệu hiệu quả hơn.