Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Unit II

Data

Types of Data

Data Pre-processing

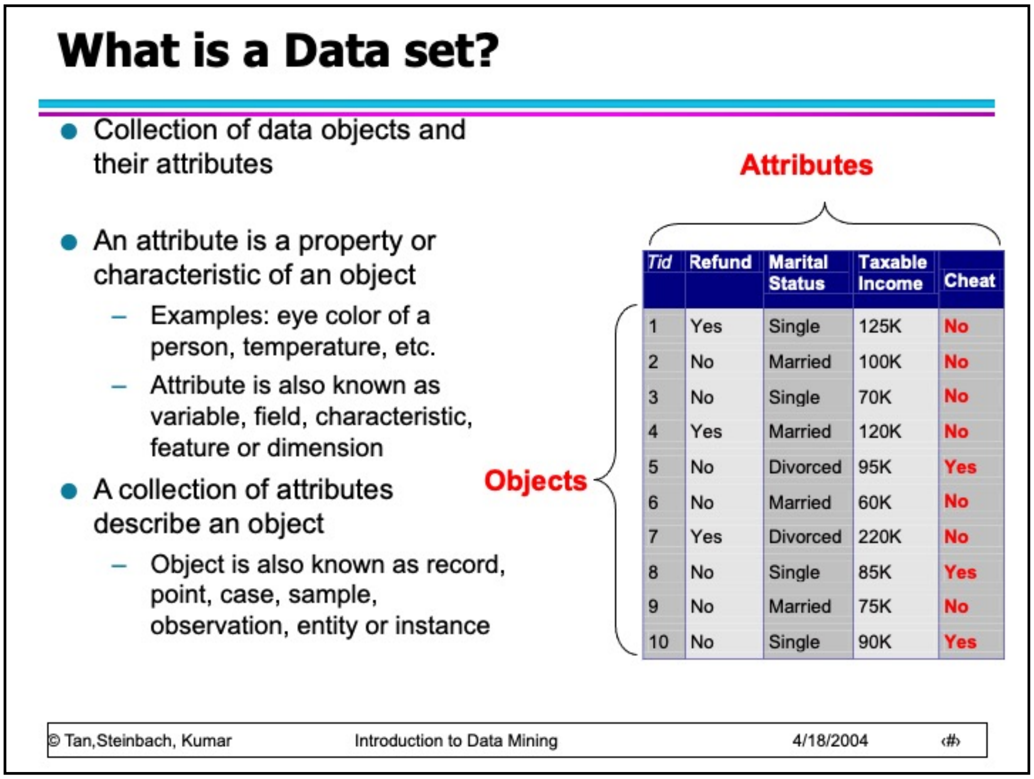
Measures of Similarity and Dissimilarity

Note : This slide is added by Dr.Vani V

Lecture 6 - Outline

- What is a data set?
- Attribute Values
- Types of Attributes
- Properties of Attribute Values
- Discrete Vs Continuous Attributes

Note : This slide is added by Dr.Vani V



What is a Data set?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, feature or dimension
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, observation, entity or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Measurement Scale

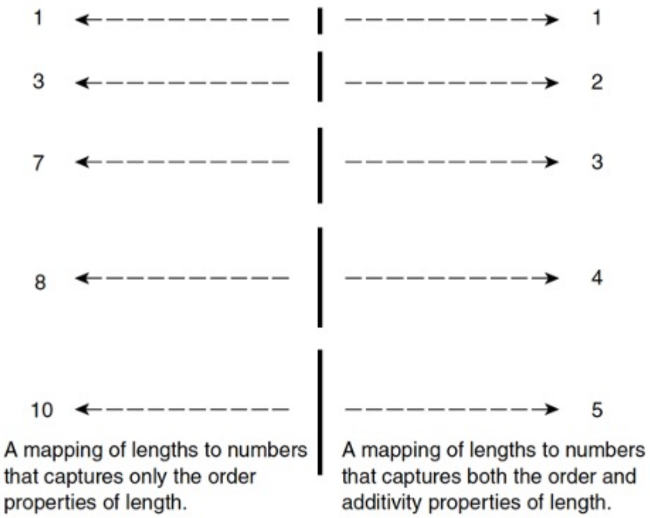
- A measurement scale is a rule (function) that associates a numerical or symbolic value with an attribute of an object.

Type of an attribute

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit, but age has a maximum and minimum value

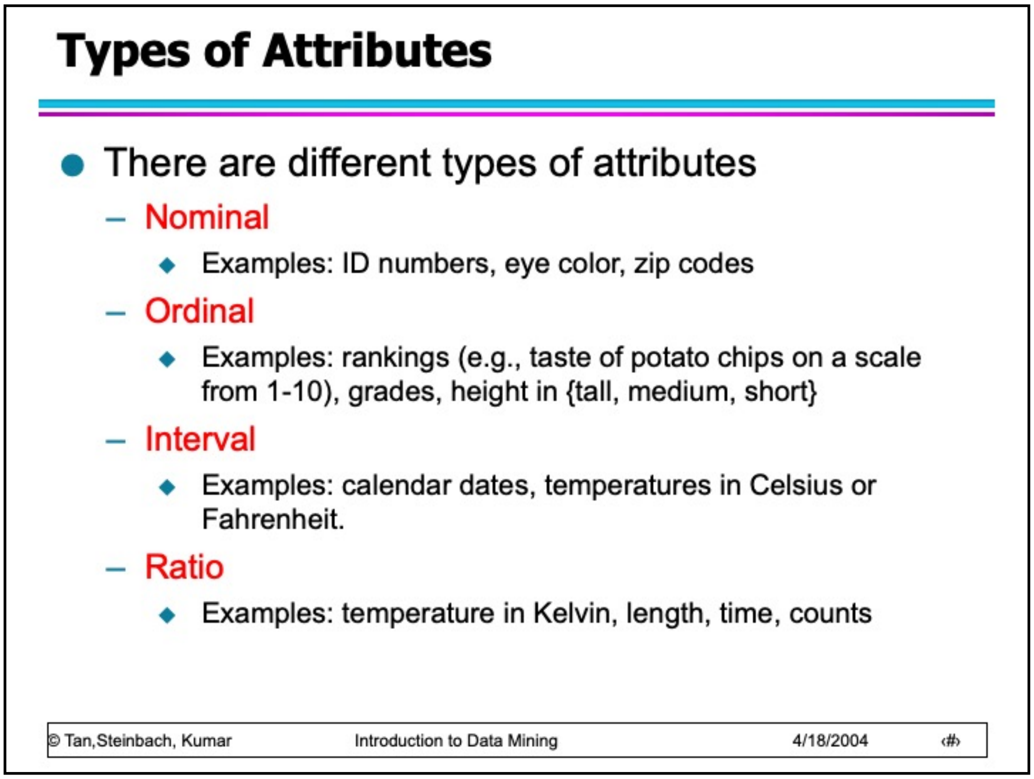
Type of an attribute



A mapping of lengths to numbers that captures only the order properties of length.

A mapping of lengths to numbers that captures both the order and additivity properties of length.

Figure 2.1. The measurement of the length of line segments on two different scales of measurement.



Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

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

Quantity (integer or real-valued)

Interval

Measured on a scale of **equal-sized units**

Values have order

E.g., *temperature in C°or F°, calendar dates*

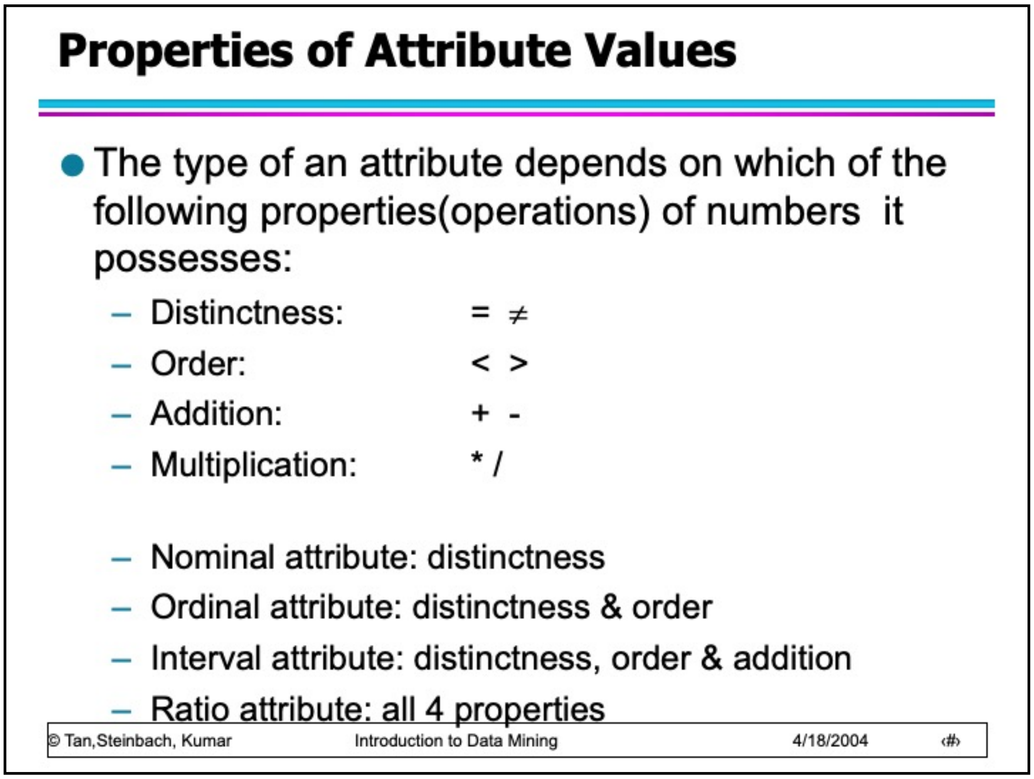
No true zero-point

Ratio

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°).

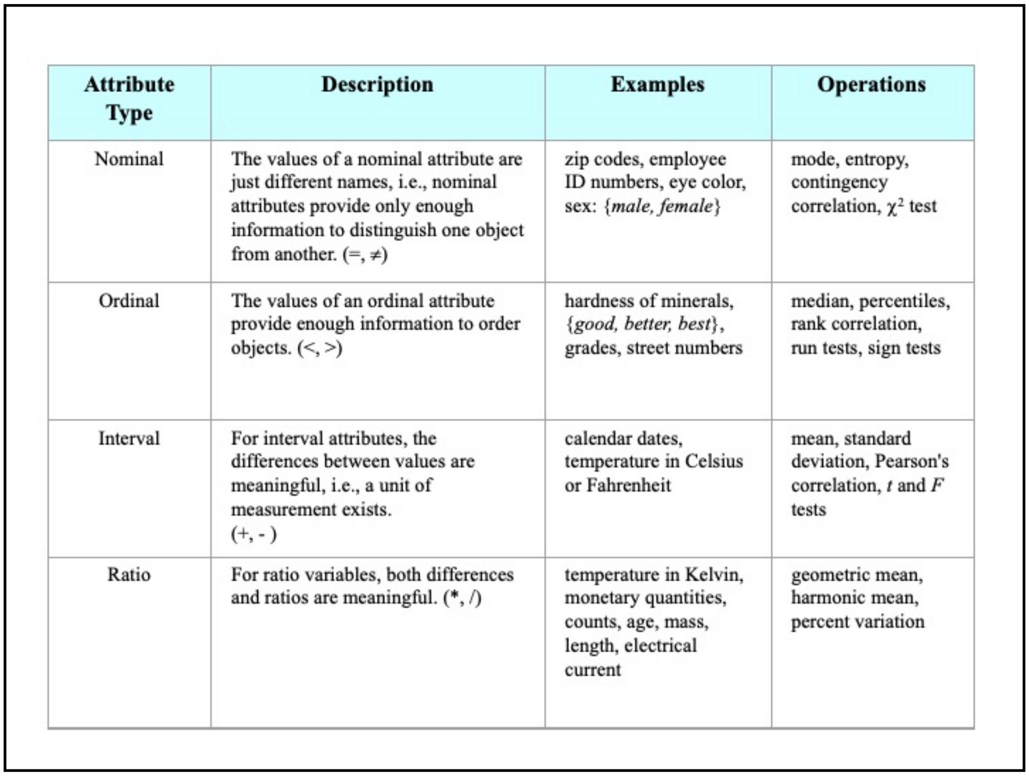
e.g., *temperature in Kelvin, length, counts, monetary* *quantities*



Properties of Attribute Values

- The type of an attribute depends on which of the following properties(operations) of numbers it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties



Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Qualitative(Categorical) : Nominal, Ordinal

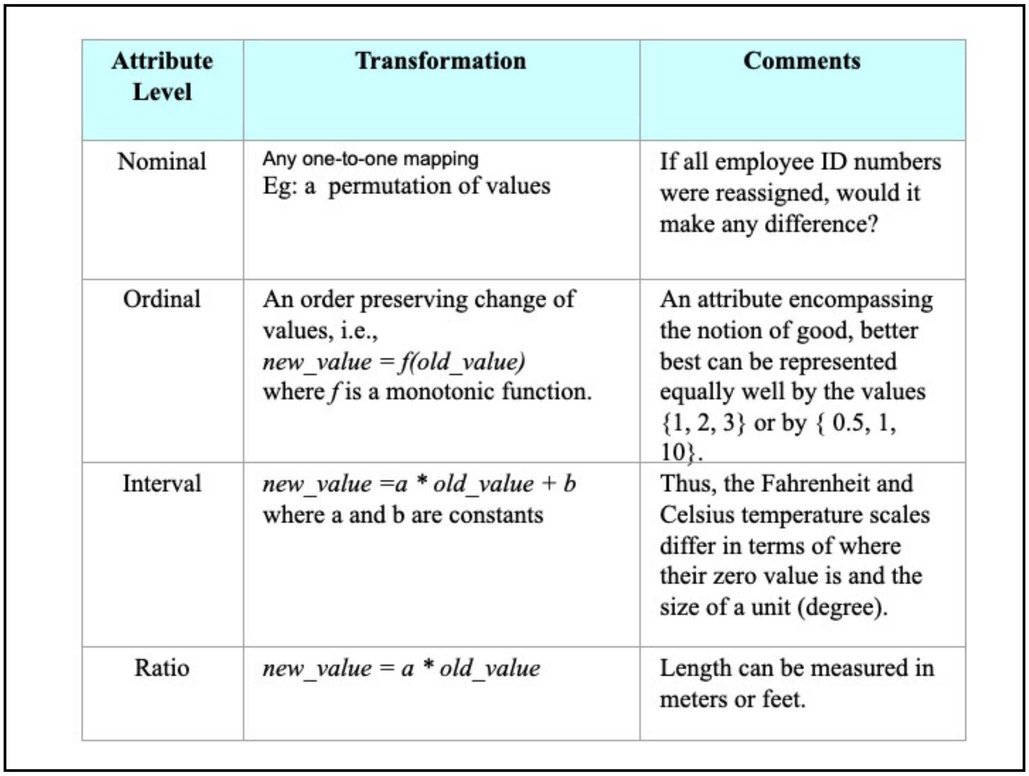
Quantitative : Interval , Ratio

An arithmetic average is the sum of a series of numbers divided by the count of that series of numbers.

The is the average of a set of products, the calculation of which is commonly used to determine the performance results of an investment or portfolio. It is technically defined as "the *nth* root product of *n* numbers." The geometric mean must be used when working with percentages, which are derived from values, while the standard arithmetic mean works with the values themselves.

The harmonic mean is best used for fractions such as rates or multiples.

The harmonic mean is a type of numerical average. It is calculated by dividing the number of observations by the reciprocal of each number in the series.



Attribute Level	Transformation	Comments
Nominal	Any one-to-one mapping Eg: a permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.