Lecture-5
Course: Applied Data Science

Attributes & Quality of Data

By
Dr. Sibarama Panigrahi
Assistant Professor, Department of Computer Sc. & Engineering
National Institute of Technology, Rourkela, Odisha, 769008, India
Mobile No.: +91-7377302566
Email: panigrahis[at]nitrkl[dot]ac[dot]in
panigrahi[dot]sibarama[at]gmail[dot]com

Outlines…

- Attributes
- Data Quality

Attributes : Defining Objects…

- **Definition:** An attribute is a property or characteristic of an object that may vary, either from one object to another or from one time to another.
  - The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature.
  - Attributes define Objects (*also called as samples, examples, instances, data points*).

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

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Attributes : Defining Objects…

- **Types of Attribute**

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	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
Numeric (Quantitative)	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 deviatio 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

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Attributes : Defining Objects…

- **Transformations defining Attribute levels:**

Attribute Type		Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values	If all employee ID numbers are reassigned, it will not 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 $\{0.5, 1, 10\}$.
Numeric (Quantitative)	Interval	$new\_value = a * old\_value + b$, a and b constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree (unit).
	Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Attributes : Defining Objects…

- **Describing Attributes by the Number of Values:**
  - Discrete:
    - A discrete attribute has a finite or countably infinite set of values.
    - Such attributes can be categorical, such as zip codes or ID numbers, or numeric, such as counts.
    - Often represented as integer variables.
    - Note: binary attributes are a special case of discrete attributes
  - Continuous:
    - A continuous attribute is one whose values are real numbers.
    - Examples include attributes such as 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.

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Attributes : Defining Objects…

- Asymmetric Attributes
  - Only presence (a non-zero attribute value) is regarded as important
    - Words present in documents
    - Items present in customer transactions

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Critiques of the attribute categorization

- Incomplete
  - Asymmetric binary
  - Cyclical
  - Multivariate
  - Partially ordered
  - Partial membership
  - Relationships between the data

- Real data is approximate and noisy
  - This can complicate recognition of the proper attribute type
  - Treating one attribute type as another may be approximately correct

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Key Messages for Attribute Types

- The types of operations you choose should be "meaningful" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four (among many possible) properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - In the end, what is meaningful can be specific to domain.

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Important Characteristics of Data

- Dimensionality (number of attributes)
  - The **dimensionality of a data set** is *the number of attributes that the objects* in the data set possess.
  - The *difficulties associated with analyzing high-dimensional data* are sometimes referred to as the **curse of dimensionality**.
  - Because of this, an important motivation in preprocessing the data is **dimensionality reduction**.
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Data Quality

- Poor data quality negatively affects many data processing efforts

- Data Science example:
  - a classification model for detecting people who are loan risks is built using poor data
    - Some credit-worthy candidates are denied loans
    - More loans are given to individuals that default

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Data Quality

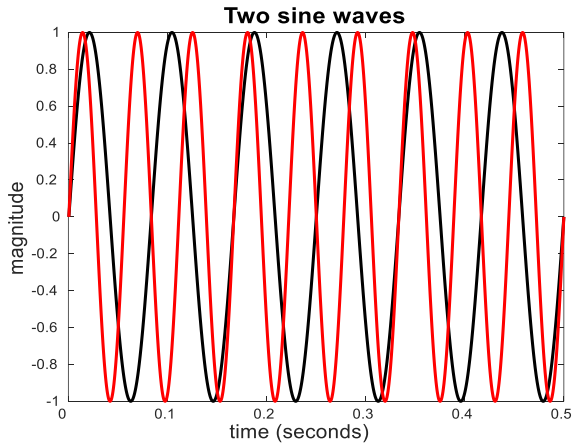
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

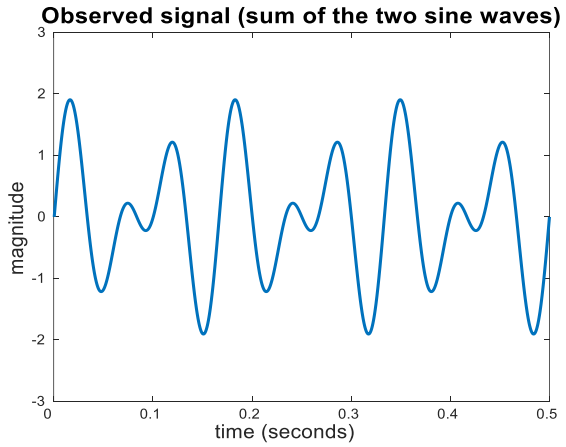
- Examples of data quality problems:
  - Noise and outliers
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

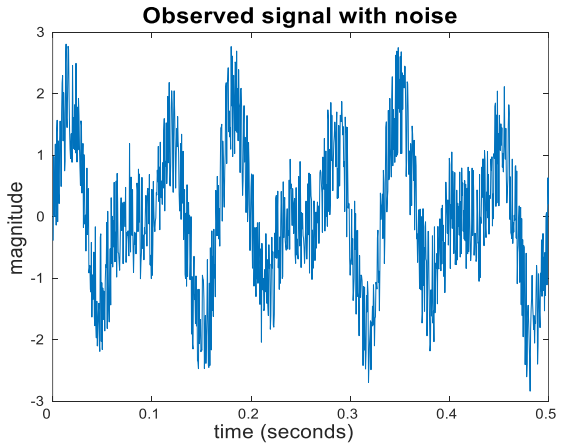
Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
  - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - The magnitude and shape of the original signal is distorted

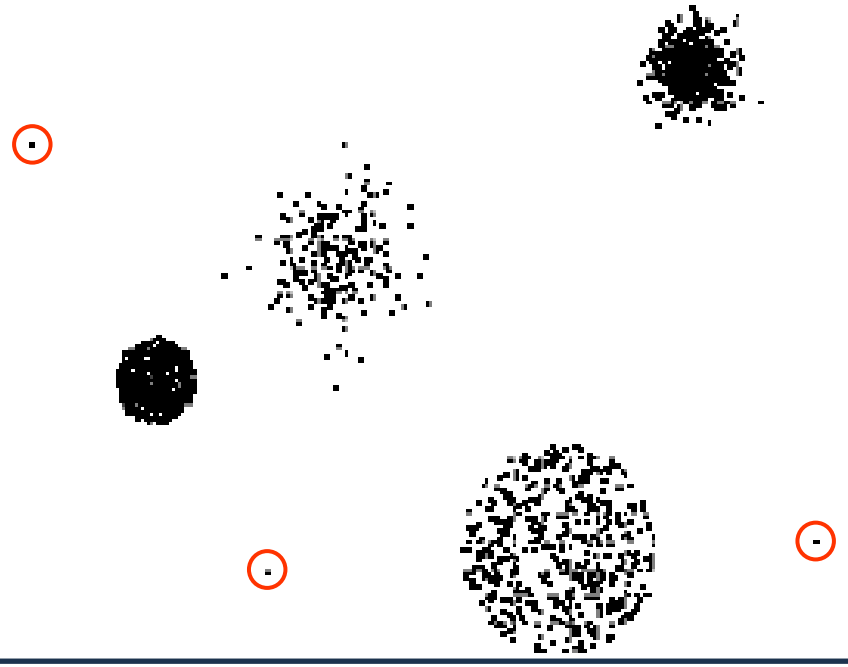






Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

Source:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

For Your Valuable Time.