Chapter 2 and 3, Data Pre-processing

Young-Rae Cho
Associate Professor
Department of Computer Science
Baylor University

Why Need Data Pre-processing?

- **Incomplete Data**
  - Missing values, or Lack of attributes of interest

- **Noisy Data**
  - Errors, or Outliers

- **Redundant Data**
  - Duplicate data, or Duplicate attributes
    - e.g., Age = "47", Birthday = "01/07/1968"

- **Inconsistent Data**
  - Containing discrepancies in format or name
    - e.g., Rating by "1, 2, 3", Rating by "A, B, C"

- **Huge Volume of Data**
Importance

- **Lower Quality Data, Lower Quality Mining Results !!**
  - Mining quality depends on data quality as well as mining techniques.

- **Majority of Data Mining**
  - Data pre-processing comprises the majority of the works for data warehousing and data mining

Major Tasks

- **Data Cleaning**
  - Fill in missing values, smooth noisy data, remove outliers, remove redundancy, and resolve inconsistency

- **Data Integration**
  - Integration of multiple databases or files

- **Data Transformation**
  - Normalization and aggregation

- **Data Reduction**
  - Reducing representation in volume with similar analytical results
  - Discretization of continuous data
Chapter 2 and 3, Data Pre-processing

- General Data Characteristics
  - Descriptive Data Summarization
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction

Data Types

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data, e.g., text documents
  - Transaction data

- Ordered Data
  - Sequential data, e.g., transaction sequences, biological sequences
  - Temporal data, e.g., time-series data
  - Spatial data, e.g., maps

- Graph
  - WWW, internet
  - Social or information networks
  - Biological networks
Attribute Types

- **Nominal**
  - e.g., ID number, profession, zip code

- **Ordinal**
  - e.g., ranking, grades, sizes

- **Binary**
  - e.g., medical test (positive or negative)

- **Interval**
  - e.g., calendar dates, temperature, height

- **Ratio**
  - e.g., population, sales

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Finite set of values
  - Sometimes, represented as integer values
  - Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Real numbers as values
  - Typically, represented as floating-point variables
  - In practice, shown as a finite number of digits
Characteristics of Data

- **Dimensionality**
  - Curse of dimensionality

- **Sparsity**
  - Lack of information

- **Resolution**
  - Patterns depending on the scale

- **Similarity**
  - Similarity measures for complex types of data

---

**CSI 4352, Introduction to Data Mining**

**Chapter 2 and 3, Data Pre-processing**

- General Data Characteristics
- Descriptive Data Summarization
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
Descriptive Data Mining

- **Motivation**
  - To better understand the properties of data distributions, e.g., central tendency, spread and variation

- **Measurements**
  - median, max, min, quantiles, outliers, etc.

- **Analysis Process**
  - Folding the measures into numeric dimensions
  - Graphic analysis on the transformed dimension space

Central Tendency Measures

- **Mean**
  - Weighted arithmetic mean:
  
  \[ \bar{x} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j} \]
  - Trimmed mean: chopping extreme values

- **Median**
  - Middle value if odd number of values
  - Average of two middle values otherwise
  - Estimation by interpolation for grouped data:
    \[ \text{median} = L_n + \left( \frac{N/2 - \sum \text{freq}_{med}}{\text{freq}_{med}} \right) \text{width} \]

- **Mode**
  - The value that occurs the most frequently in the data
  - Unimodal, bimodal, trimodal distribution
Central Tendency in Skewed Data

- **Symmetric Data**

- **Skewed Data**

Data Dispersion Measures

- **Quartiles and Outliers**
  - Quartiles: $Q_1$ (25th percentile), $Q_3$ (75th percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Outliers: data with extreme low and high values
    usually, values lower/higher than $Q_1 - 1.5 \times IQR / Q_3 + 1.5 \times IQR$

- **Variance and Standard Deviation**
  - $\sigma^2$, $\sigma$ in population:
    \[
    \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2
    \]
  - $s^2$, $s$ by sampling:
    \[
    s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right]
    \]

**Degree of Freedom**: # independent pieces of information

(= # independent measurement - # parameters)
Graphic Analysis

- **Boxplot**
  - Display of five-number summary

- **Histogram**
  - Display of tabulated frequencies

- **Quantile-Quantile (Q-Q) Plot**
  - Description of the relationship between two univariate distributions

- **Scatter Plot**
  - Description of the relationship between two attributes of a bivariate distribution

Boxplot Analysis

- **Five-number summary of a Distribution**
  - Minimum / \( Q_1 \) / Median / \( Q_3 \) / Maximum

- **Boxplot**
  - Represented as a box
  - The bottom of the box is \( Q_1 \)
  - The top of the box is \( Q_3 \)
  - The median is marked by a line
  - Whiskers: two lines outside of the box extend to minimum and maximum
Histogram Analysis

- **Histogram**
  - Univariate graphic method
  - Represented as a set of bars reflecting the frequencies of the discrete values
  - Grouping data values into classes if they are continuous

- **Boxplot vs. Histogram**
  - Often, histogram gives more information than boxplot

![Histogram Example](image1)

Quantile Plot Analysis

- **Quantile Plot**
  - Plots quantile information of the data (sorted in an ascending order)
  - Displays all the data

- **Q-Q (Quantile-Quantile) Plot**
  - Plots the quantiles of one univariate distribution against the quantiles of the other
  - Describes the relationship between two distributions

![Q-Q Plot Example](image2)
Scatter Plot Analysis

- **Scatter Plot**
  - Displays the points of bivariate data
  - Describes the relationship between two attributes (variables)

![Scatter Plot Diagram]

- **Positively correlated data**
- **Negatively correlated data**
- **Clusters**
- **Patterns**
- **Outliers**

---

CS 4352, Introduction to Data Mining

**Chapter 2 and 3, Data Pre-processing**

- General Data Characteristics
- Descriptive Data Summarization
  - **Data Cleaning**
- Data Integration
- Data Transformation
- Data Reduction