1.2 Data Preparation

Where the majority of the time is spent on any data mining project

- importing, manipulating, cleaning, transforming, augmenting

Data

- Collection of data objects (cases) and their attributes (features)
  - **Attribute**: property/characteristic of an object
  - **Object**: collection of attributes
- Can be **structured** (data table) or **non-structured** (text)
- Can have **non-dependency** or **dependency** (time, space)

Types of Data Sets

- **Nondependency-oriented data**: no dependencies between cases
- **Dependency-oriented data**: implicit/explicit relationships between cases

Types of Attributes

Categorical/Qualitative Attributes

- **Nominal**: no relationship between values
- **Ordinal**: order between the values, but no mathematical operation can be performed on them

Numeric/Quantitative Attributes

- **Discrete**: finite/countably infinite set of values for which differences are meaningful
- **Continuous**: infinite set of values that represent the absolute numbers

Important Characteristics

- Dimensionality
- Sparsity
- Resolution
- Size

Data Preparation

Data analysis tasks use source data sets stored in tabular format

Data Wrangling

Transform and map data from one "raw" data form into another format appropriate for analytics

- **Goal**: attain quality and useful data

Steps

1. Discovering
2. Structuring
3. Cleaning
4. Enriching
5. Validating
6. Publishing

Data Quality

- Raw → values may be missing, inconsistent across different data sources, erroneous → poor data quality

Data Quality Problems

- Noise and outliers
- Missing values
- Duplicate data
- Inconsistent/incorrect data

Noise

- Irrelevant or useless information
- **Possible causes**: incorrect/distorted measurements, proper variability of the domain

Outliers

- Data objects with characteristics that are considerably different from most of the other objects in the data sets
- **Cases**: outlier = noise or outliers = goal

Missing Values

- **Missing Completely at Random (MCAR)**: missing value is independent of observed and unobserved data; nothing systematic about it
    - e.g.: a lab value because a lab sample was processes improperly

- **Missing at Random (MAR)**: missing value is related to observed data, not to unobserved data; may be something systematic about it
    - e.g.: missing income value may depend on the age
- **Missing Not a Random (MNAR)**: missing value is related to unobserved data of the variable itself; informative/non-ignorable missingness
    - e.g.: a person did not enter his/weight in a survey

Solutions

- **Remove**: critical if there are meany observations with missing values
- **Ignore**
- **Make estimates**: imputation
    - the most common value of the attribute; based on other attributes; more sophisticated methods
    - might introduce bias in data and affect the results

Duplicates

- Major issues when merging data from heterogeneous source

Inconsistent/Incorrect Data

- The hardest type of data quality issues to detect
- Depends on expert domain knowledge