Data Preprocessing

- Data in the real world is not clean
- No quality data, no quality mining results!
- Quality decisions must be based on quality data

Data Cleaning

Fill in missing values, smooth noisy data, identify or remove outliers,and resolve inconsistencies

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

Missing Data: Main Causes

- Data is not always available
- Equipment malfunction
- Inconsistent with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Certain data may not be considered important at the time of entry
- Not register history or changes of the data

How to Handle Missing Data?

- Ignore the tuple with missing class label
- Fill in the missing value manually
- Fill in it automatically witha global constant, the attribute mean, the attribute mean for all samples belonging to the same class: smarter, the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data: Main Causes

- Noise: random error or variance in a measured variable
- Faulty data collection instruments, data entry problems, data transmission problems, technology limitation, inconsistency in naming convention
- Duplicate records, incomplete data, inconsistent data

How to Handle Noisy Data?

- Binning: first sort data and partition into (equal-frequency) bins, then one can smooth by bin means, smooth by bin median,smooth by bin boundaries, etc.
- Regression: smooth by fitting the data into regression functions
- Clustering: detect and remove outliers
- Combined computer and human inspection□ detect suspicious values and check by human (e.g., deal withpossible outliers)

Data Integration

Integration of multiple databases, data cubes, or files

- **Data integration:** Combines data from multiple sources
- **Schema integration:** e.g., A.cust-id ≡ B.cust-#: Integrate metadata from different sources
- **Entity identification problem:** Identify real world entities from multiple data sources, e.g., Bill Clinton =William Clinton, Detecting and resolving data value conflicts, For the same real world entity, attribute values from different sources aredifferent, Possible reasons: different representations, different scales, e.g., metric vs.British units

Handling Redundancy in Data Integration

- Redundant data occur often when the integration of multiple databases: **Object identification:** The same attribute or object may have different names in different databases, **Derivable data:** One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

Normalization and aggregation

- **Smoothing:** remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range, min-max normalization, z-score normalization, normalization by decimal scaling
- Attribute/feature construction, New attributes constructed from the given ones

Data Reduction

Obtains reduced representation in volume but produces the same or similar analytical results

- **Why data reduction?**: A database/data warehouse may store terabytes of data, Complex data analysis/mining may take a very long time to run on thecomplete data set
- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- **Data reduction strategies**: Dimensionality reduction, e.g., remove unimportant attributes, Data Compression: lossy and lossless, Numerosity reduction, e.g., fit data into models, Discretization and concept hierarchy generation

Data Discretization

Part of data reduction but with particular importance, especially for numerical data

- Binning
- Histogram analysis
- Clustering analysis
- Entropy-based discretization
- Interval merge by χ^2 analysis
- Segmentation by natural partitioning