Search and Data Mining: Techniques

Data Preprocessing

Anya Yarygina

Boris Novikov

Introduction

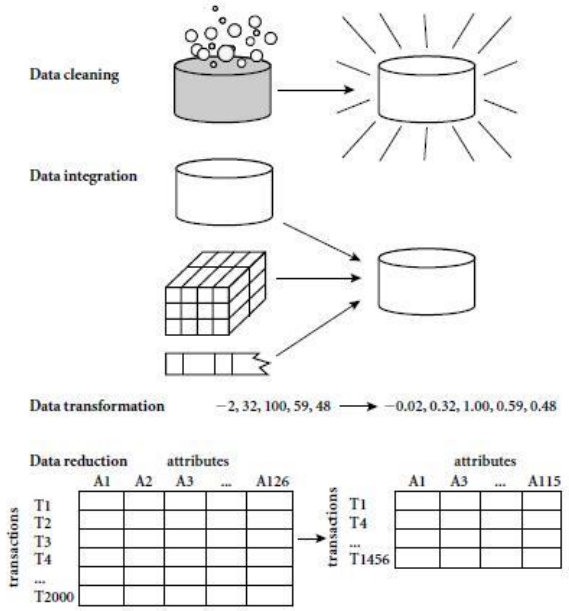
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Data discretization

Reference

Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, 2nd Edition. The Morgan Kaufmann Series in Data Management Systems, 2006

Why preprocess the data?

- Data
  - Incomplete
    Lacking attribute values or certain attributes of interest; containing only aggregate data
  - Noisy
    Containing errors or outliers
  - Inconsistent



- Data cleaning
  - Fill missing values
  - Smooth noisy data
  - Identify or remove outliers
  - Resolve inconsistencies
- Data Integration
  - Redundant data
- Data transformation
  - Normalization
  - Aggregation
- Data reduction
  - Data aggregation
  - Attribute subset selection
  - Dimensionality reduction
  - Numerosity reduction

Descriptive data summarization

- Identify typical properties of your data and highlight which data values should be treated as noise or outliers
  - Central tendency
  - Dispersion
  - Graphic display

Measuring the central tendency

- Mean
- Weighted arithmetic mean
- Trimmed mean
- Median
- Mode (value that occurs most frequently in the set)
- Midrange (average of the largest and smallest values in data set)

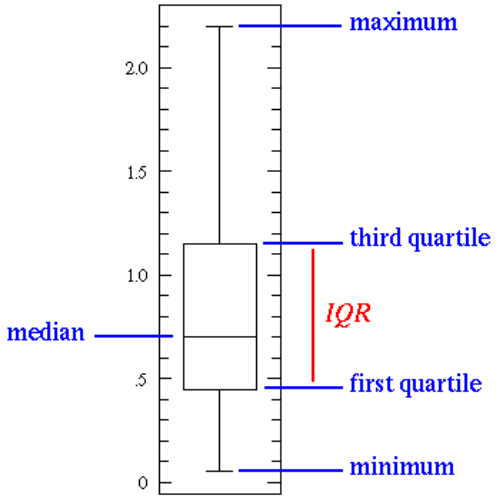
- Distributive measure
- Algebraic measure
- Holistic measure

Measuring the dispersion of data

- Range
- Five-number summary (based on quartiles)
- Interquartile range
- Standard deviation

Range, Quartiles, Outliers, and Boxplots

- Range
- Interquartile range (IQR)
- Outliers (1.5*IQR above the third quartile)
- Five-number summary
  - Boxplot
- K-th percentile
  - Median
  - Quartiles

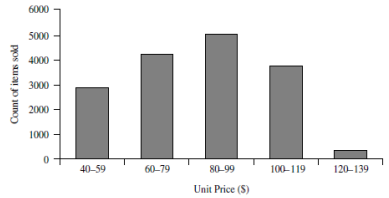


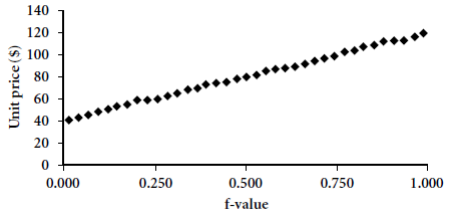
Variance and standard deviation

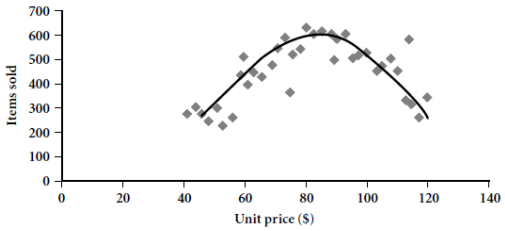
- Variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$
- Standard deviation
  - Spread about the mean
  - Zero standard deviation
- Algebraic measures

Graphic displays of basic descriptive data summaries

- Histograms (frequency histograms)
- Quantile plot
- Scatter plot
- Loess curve







Data cleaning

- Fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data
  - Missing values
  - Noisy data

Missing values

- Ignore the tuple
- Fill the missing value manually
- Use the global constant to fill the missing value
- Use the attribute mean to fill the missing value
- Use the attribute mean for all samples belonging to the same class as the given tuple
- Use the most probable value to fill in the missing value

Noisy data

- Noise is a random error or variance in a measure variable
  - Binning
    - Smoothing by bin means
    - Soothing by bin medians
    - Smoothing by bin boundaries
  - Regression
  - Clustering

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data cleaning as a process

- Discrepancy detection
  - Data scrubbing tools
  - Data auditing tools
- Data transformation
  - Data migration tools
  - ETL tools

Data integration and transformation

- Merging of data from multiple data stores
- Data should be transformed into forms appropriate for mining

Data integration

- Combines data from multiple data sources (databases, data cubes, flat files) into a coherent data store
  - Schema integration and object matching
    - Entity identification problem
  - Redundancy
    - Correlation analysis
  - Detection and resolution of data value conflicts

Correlation Analysis

- Correlation coefficient $r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$
  - Independence, positively and negatively correlated
  - Correlation does not imply causality
- Chi-square test
  - Contingency table $e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$
  - Pearson chi-square statistic $\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

Data transformation

- Data are transformed or consolidated into forms appropriate for mining
  - Smoothing (remove noise from data)
    - Binning
    - Regression
    - Clustering
  - Aggregation
  - Generalization (concept hierarchies)
  - Normalization
    - Min-max normalization
    - Z-score normalization
    - Normalization by decimal scaling
  - Attribute construction

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

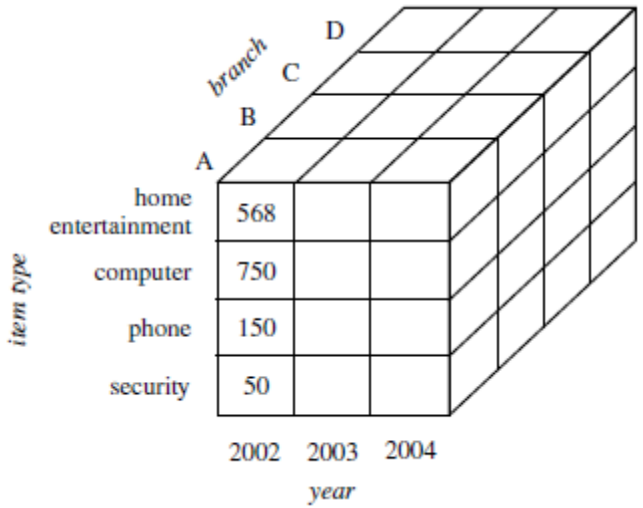
$$v' = \frac{v}{10^j}$$

Data reduction

- Obtain reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data
  - Data cube aggregation
  - Attribute subset selection
  - Dimensionality reduction
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Data cube aggregation

- Data cubes store multidimensional aggregated information
  - Item type
  - Branch
  - Year
  - Total sales



Attribute subset selection

- Reduce data set size by removing irrelevant or redundant attributes
- Find a minimum set of attributes such that resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes

- Stepwise forward selection
- Stepwise backward elimination
- Combination of forward selection and backward elimination
- Decision tree induction

Dimensionality reduction

- Reduced or “compressed” representation of the original data
  - Discrete wavelet transforms
  - Principal components analysis

Numerosity reduction

- Reduce the data volume by choosing alternative, |smaller| forms of data representation
  - Parametric methods
    - Regression $y = wx + b$
    - Log-linear models
  - Nonparametric methods
    - Histograms
      - Equal-width
      - Equal-frequency
      - V-optimal (least variance as weighted sum of the original values that each bucket represents, where bucket weight is equal to the number o values in the bucket)
      - MaxDiff
    - Clustering
    - Sampling
      - Simple random sample without replacement
      - Simple random sample with replacement
      - Cluster sample
      - Stratified sample

Data discretization and concept hierarchy generation

- Supervised and unsupervised discretization
- Top-down and bottom-up discretization
- Concept hierarchy

Data discretization and concept hierarchy generation for numerical data

- Binning
- Histogram analysis
- Entropy-based discretization
- Interval merging by chi-squared analysis
- Cluster analysis
- Discretization by intuitive partitioning

Concept hierarchy generation for categorical data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
- Specification of a portion of a hierarchy by explicit data grouping
- Specification of a set of attributes, but not of their partial ordering
- Specification of only a partial set of attributes

Outline

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Data discretization