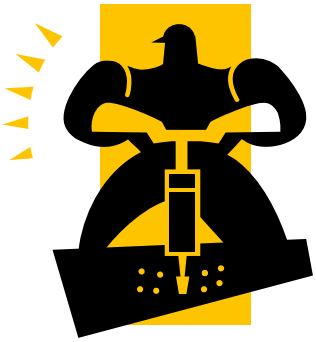


Data Mining -- Data Preparation

Instructor: Jen-Wei Huang

Office: 92528 in the EE building
jwhuang@mail.ncku

Data Types

- Texts
  - Relational records, Transactions
  - Documents
  - Logs
- Sequences
  - Time series
  - Videos
  - Biomedical sequence
- Special structure
  - Web pages, forums
  - Graphs, social networks, interactions
  - Multimedia, images, videos
  - Spatial data, maps

Characteristics of Data

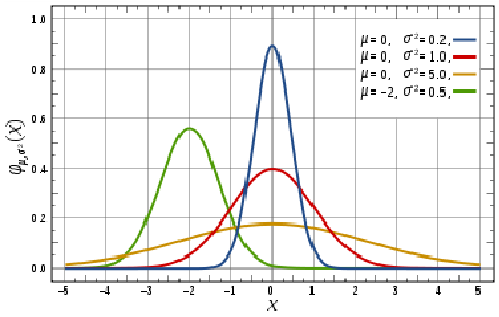
- Dimensionality
  - Curse of dimensionality
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion
- Similarity
  - Numerical or categorical data

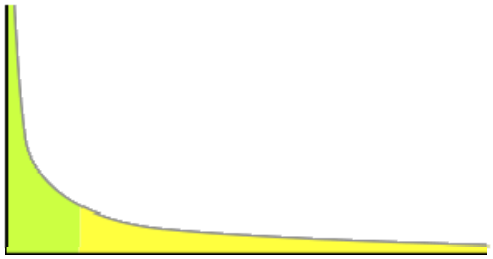
Measurements of Data

- Minimum, maximum
- Mean, median, mode
- Variance, standard deviation
- Quartiles
  - min, Q1, mean, Q3, max
  - IQR: inter-quartile range = Q3-Q1
- Skewness
  - http://en.wikipedia.org/wiki/Skewness
- Outliers
- Correlation coefficient
  - http://en.wikipedia.org/wiki/Correlation_coefficient

Data Distributions

- Normal distribution
  - http://en.wikipedia.org/wiki/Normal_distribution
- Power law distribution
  - http://en.wikipedia.org/wiki/Power_law





Preprocessing

- Data cleaning
  - Incomplete data cleaning
  - Noise/outliers reduction
- Data Integration
  - Entity resolution
- Data reduction
  - Sampling
  - Dimension reduction
  - Compression
- Data transformation
  - Domain transformation
  - Discretization

Data Warehousing

- "subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

  by W. H. Inmon

- Types:
  - Plain texts
  - Records
  - Vectors
  - Relations
  - Databases
  - Data cubes and OLAP(online analytical processing)

References

- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997

References

- Slides from Prof. J.-W. Han, UIUC
- Slides from Prof. W.-Z. Peng, NCTU