DATA WAREHOUSING & MINING

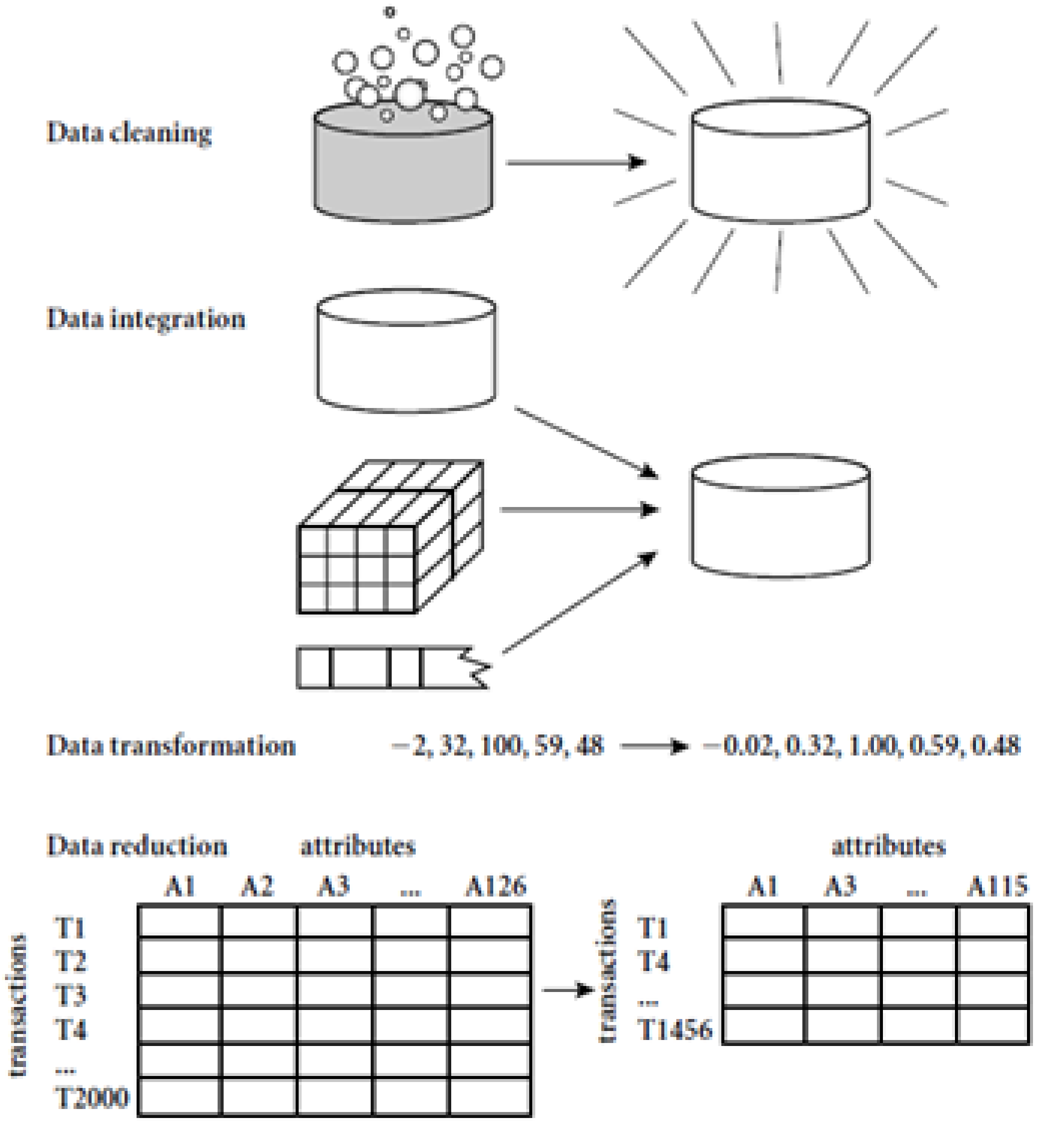
Unit-1: Introduction – Data Preprocessing

Parameshwar R Hegde, PhD

Why Data Preprocessing?

- Collection instruments used may be faulty
- Human or computer errors occurring at data entry
- Disguised missing data - Users may purposely submit incorrect data values for mandatory fields
- Limited buffer size for coordinating synchronized data transfer and consumption
- Two other factors affecting data quality are believability and interpretability

Data Preprocessing Forms



Data cleaning

Data integration

Data transformation −2, 32, 100, 59, 48 → −0.02, 0.32, 1.00, 0.59, 0.48

Data reduction

attributes

transactions	A1	A2	A3	...	A126
T1					
T2					
T3					
T4					
...					
T2000					

→

attributes

transactions	A1	A3	...	A115
T1				
T4				
...				
T1456				

Data Cleaning and Integration

- **Data integration - merges data from multiple sources into a coherent data store such as a data warehouse**

- **Data cleaning - Remove noise and correct inconsistencies in data**

- **“Clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies**
  - ➢ Ex., During data integration, data attributes representing a given concept may have different names in different databases
    - ▪ *customer_salary* in one data store and *cust_salary* in another

- **Can be performed to detect and remove redundancies**

Data Cleaning and Integration

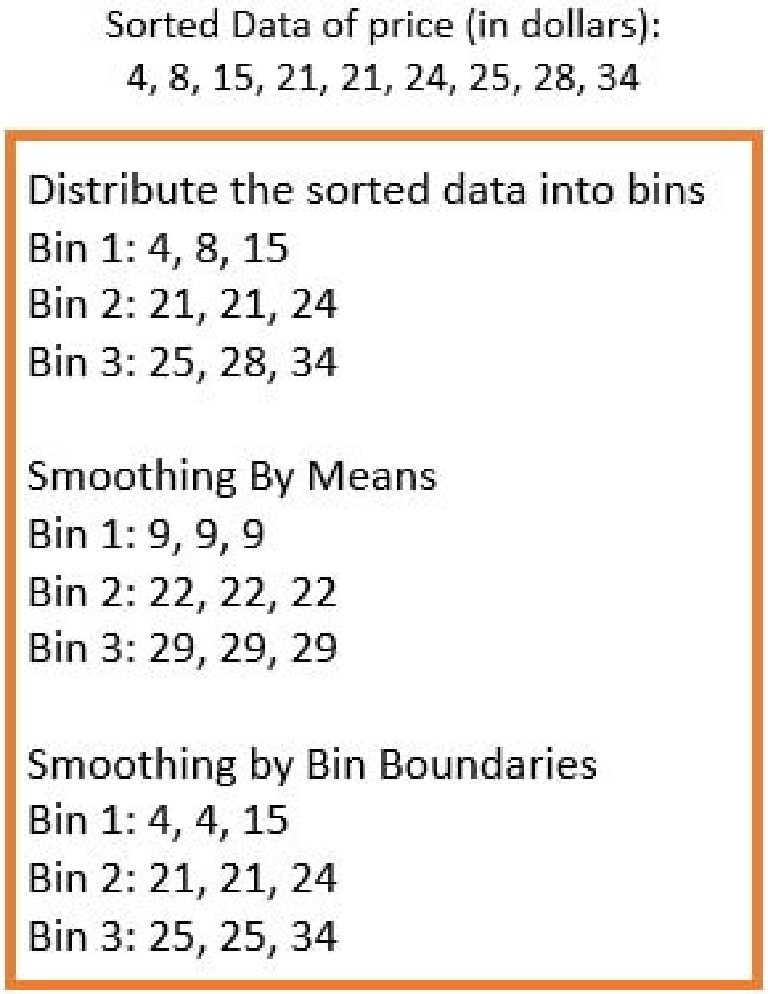
- **Missing values -** Filling in the missing values for this attribute
  1. **Ignore the tuple:** Done when the class label is missing
     - Not very effective, unless the tuple contains several attributes with missing values
  2. **Fill in the missing value manually:** Is time consuming and may not be feasible given a large data set with many missing values
  3. **Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant such as a label like "Unknown" or $-\infty$
  4. **Use a measure of central tendency for the attribute:** Fill in the missing value by measures of central tendency, which indicate the "middle" value of a data distribution
     - For Symmetric data distributions, the mean can be used, while skewed data distribution should employ the median

Data Cleaning

- **Missing values -** Filling in the missing values for this attribute
  5. **Use the attribute mean or median for all samples belonging to the same class as the given tuple**
  6. **Use the most probable value to fill in the missing value:** Predict the missing value and fill it
     - Determined with regression, inference-based tools using a Bayesian formalism, or decision tree

Data Cleaning

- **Noisy data -** A random error or variance in a measured variable
  1. **Binning:** Smooth a sorted data value by consulting its “neighborhood,”
     - Sorted values are distributed into a number of buckets or bins



Sorted Data of price (in dollars):
4, 8, 15, 21, 21, 24, 25, 28, 34

Distribute the sorted data into bins
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing By 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

- **Noisy data -** A random error or variance in a measured variable
  2. **Regression:** Linear regression - Finding the “best” line to fit two attributes so that one attribute can be used to predict the other
  3. **Outlier analysis:** Outliers may be detected by clustering
     - Ex., Similar values are organized into groups, or clusters

Data Integration

- **Merging of data from multiple data stores**

- **Careful integration can help**
  - ➢ Reduce and avoid redundancies and inconsistencies
  - ➢ Improve the accuracy and speed of the subsequent data mining process

Data Integration

- **Might phase the problems like**
  1. **Entity Identification Problem:**
     - Multiple sources may include multiple databases, data cubes, or flat files
     - Schema integration and object matching can be tricky
       Ex., customer_id in one database and cust_number in another database
       - Metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling this
     - Matching attributes from one database to another during integration, special attention must be paid to the structure of the data
       Ex., Discount may be applied to the order, whereas in another system it is applied to each individual line item within the order

Data Integration

- **Might phase the problems like**
  2. **Redundancy and Correlation Analysis:**
     - An attribute may be redundant if it can be “derived” from another attribute or set of attributes

       Ex., Annual revenue derived from No. of sales
     - For nominal data, correlation is relationship between two attributes

       Ex., Relationship between A and B, can be discovered by a χ2 (chi-square) test
     - Correlation Coefficient for Numeric Data; evaluate the correlation between two attributes

       Ex., A and B, by computing the correlation coefficient (known as Pearson's product moment coefficient)

Data Integration

- **Might phase the problems like**

  3. **Tuple Duplication:**
     - Detecting redundancies between attributes, duplication should also be detected at the tuple level

       Ex., In purchase order database; purchaser's name appearing with different addresses
     - Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences

Data Integration

- **Might phase the problems like**
  4. **Data value conflict detection and resolution:**
     - Detection and resolution of data value conflicts

       Eg., In hotel chain database; price of rooms in different cities may involve not only different currencies but also different services (Eg., free breakfast) and taxes
     - May be due to differences in representation, scaling, or encoding

Data Reduction

- **Reduce data size by, for instance, aggregating, eliminating redundant features or clustering**

- **Strategies includes,**
  1. **Dimensionality reduction -** Data encoding schemes are applied so as to obtain a reduced or “compressed” representation of the original data
     - ➢ Ex., Data compression techniques – Principle Components Analysis (PCA)

       Attribute subset selection – Removing irrelevant attributes

       Attribute construction - Small set of more useful attributes is derived from the original set

Data Reduction

- **Strategies includes,**

2. **Numerosity reduction -** Data are replaced by alternative, smaller representations using 2 models
   - ➢ Parametric models – Ex., Regression or log-linear models
   - ➢ Nonparametric models – Ex., Histograms, clusters, sampling or data aggregation

3. **Data compression -** Obtain a reduced or "compressed" representation of the original data
   - ➢ Lossless reduction - Reconstructed from the compressed data without any information loss
   - ➢ Lossy reduction - Reconstruct only an approximation of the original data

Data Reduction

- **Dimentionality reduction**
  1. **Principal components analysis(PCA) -** also called the Karhunen-Loeve, or K-L, method
     - Ex., k n-dimensional orthogonal vectors that can best be used to represent the data, where k ≤ n
     - The input data are normalized, so that each attribute falls within the same range
     - Principal components essentially serve as a new set of axes for the data

Data Reduction

- **Dimentionality reduction**

  2. **Attribute Subset Selection:** Do not select the attributes which may be irrelevant to the mining task or redundant
     - ➢ Ex., For student result analysis, Student address, phone number are irrelevant
     - ➢ The “best” (and “worst”) attributes are typically determined using tests of statistical significance,

       I. **Stepwise forward selection:** Best of the original attributes is determined and added to the empty set of attribute

       II. **Stepwise backward elimination:** The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set

Data Reduction

- **Dimentionality reduction**
  2. **Attribute Subset Selection:** Do not select the attributes which may be irrelevant to the mining task or redundant
     - III. **Combination of forward selection and backward elimination:** The procedure selects the best attribute and removes the worst from among the remaining attributes.
     - IV. **Decision tree induction:** Decision tree algorithms (e.g., ID3, C4.5, and CART) were originally intended for classification
       - Constructs a flowchart-like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction

Data Reduction

- **Dimentionality reduction**

3. **Histograms:** Binning to approximate data distributions and are a popular form of data reduction.
   - ➢ **Equal-width:** The width of each bucket range is uniform
   - ➢ **Equal-frequency (or equal-depth):** Each bucket contains roughly the same number of contiguous data samples

4. **Clustering:** Partition the objects into groups, or clusters, so that objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters

5. **Sampling:** Allows a large data set to be represented by a much smaller random data sample (or subset)

Data Transformation Strategies Overview

- **Data transform -** consolidated into forms appropriate for mining
  - ➢ **Smoothing:** Remove noise from the data. Techniques include binning, regression, and clustering
  - ➢ **Attribute construction (or feature construction):** New attributes are constructed and added from the given set of attributes to help the mining process
  - ➢ **Aggregation:** Summary or aggregation operations are applied to the data

    Eg., Daily sales data may be aggregated so as to compute monthly and annual total amounts.
  - ➢ **Discretization:** Raw values of a numeric attribute (Eg., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (Eg., youth, adult, senior)

Data Transformation Strategies Overview

- **Data discretization -** can be categorized based on how the discretization is performed
  - Whether it uses class information or which direction it proceeds (i.e., top-down vs. bottom-up)
  - **Supervised discretization:** Discretization process uses class information
  - **Unsupervised discretization:** Discretization process do not use class information
  - **Top-down discretization or splitting:** First finding one or a few points (called *split points* or *cut points*) to split the entire attribute range, and then repeats this recursively on the resulting intervals
  - **Bottom-up discretization or merging:** Considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals

Self Study

- **Data Transformation and Discretization by**
  - Normalization
  - Binning
  - Histogram analysis
  - Clustering