# Chapter 2

# Data Preprocessing

# What is an attribute?

- An attribute is a property or characteristic of an object. Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object. Object is also known as record, point, case, sample, entity, or instance.
- Attribute values are numbers or symbols assigned to an attribute
- Same attribute can be mapped to different attribute values. Example: height can be measured in feet or meters.
- Different attributes can be mapped to the same set of values. Example: Attribute values for ID and age are integers but properties of attribute values can be different. ID has no limit but age has a maximum and minimum value.

# Types of Attributes ( Approach 1)

| Attribute Type | Description | Examples |
|---|---|---|
| Nominal | The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ‾) | zip codes, employee ID numbers, eye color. |
| Ordinal | The values of an ordinal attribute provide enough information to order objects. (<, >) | hardness of minerals, {*good*, *better*, *best*}, grades, street numbers |
| Interval | For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists.<br>(+, - ) | calendar dates, temperature in Celsius or Fahrenheit |
| Ratio | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current |

# Types of Attribute (Approach 2)

- **Discrete Attribute**
  - Has only a finite or countable infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

# Types of Attribute (Approach 3)

- **Character**
  - values are represented in forms of character or set of characters (string).
- **Number**
  - values are represented in forms of number. Number may be in form of whole number, decimal number.

# Types of data sets

**Record**

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

**Data Matrix**

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

| Projection of x Load | Projection of y load | Distance | Load | Thickness |
|---|---|---|---|---|
| 10.23 | 5.27 | 15.22 | 2.7 | 1.2 |
| 12.65 | 6.25 | 16.22 | 2.2 | 1.1 |

# Types of data sets

## Document Data

- Each document becomes a `term' vector, each term is a component (attribute) of the vector, the value of each component is the number of times the corresponding term occurs in the document

| | team | coach | Play | ball | score | game | wi n | lost | timeout | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

# Types of data sets

**Transaction Data**

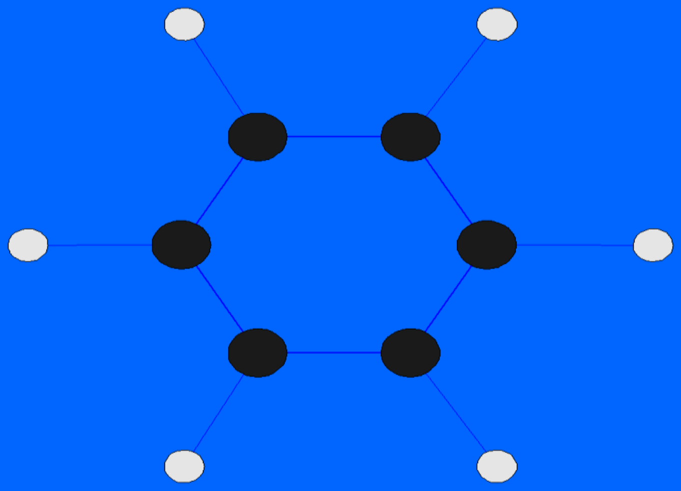
- A special type of record data, where each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

# Types of data sets

## Graph

- Contains notes and connecting vertices



# Types of data sets

**Ordered**

- Has Sequences of transactions

- **Spatial Data**
  - Spatial data, also known as geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system.
- Temporal Data
  - A temporal data denotes the evolution of an object characteristic over a period of time. Eg d=f(t).
- Sequential Data
  - Data arranged in sequence.

# Important Characteristics of Structured Data

## Dimensionality

- A Data Dimension is a set of data attributes pertaining to something of interest to a business. Dimensions are things like "customers", "products", "stores" and "time".
  - **Curse of Dimensionality**
    - When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
    - Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful
  - ***Purpose:**
    - Avoid curse of dimensionality
    - Reduce amount of time and memory required by data mining algorithms
    - Allow data to be more easily visualized
    - May help to eliminate irrelevant features or reduce noise
  - ***Techniques**
    - Principle Component Analysis
    - Singular Value Decomposition
    - Others: supervised and non-linear techniques

# Dimensionality Reduction:

## PCA

- Goal is to find a projection that captures the largest amount of variation in data.
- Find the eigenvectors of the covariance matrix.
- The eigenvectors define the new space.
- Construct a neighborhoods graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances

## Feature Subset Selection

- Another way to reduce dimensionality of data.
- Redundant features.
- Duplicate much or all of the information contained in one or more other attributes.
- Example: purchase price of a product and the amount of sales tax paid.

## Irrelevant features

- Contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

# Dimensionality Reduction:

**Techniques:**

- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches:
  - Features are selected before data mining algorithm is run

**Wrapper approaches:**

- Use the data mining algorithm as a black box to find best subset of attributes.

**Feature Creation**

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.
- Three general methodologies:
  - Feature Extraction: domain-specific
  - Mapping Data to New Space
  - Feature Construction: combining features

# Sparsity and Density

- Sparsity and density are terms used to describe the percentage of cells in a database table that are not populated and populated, respectively. The sum of the sparsity and density should equal 100.
- Many of the cell combinations might not make sense or the data for them might be missing.
- In the relational world storage of such data is not a problem: we only keep whatever there is. If we want to keep closer to our multidimensional view of the world, we face a dilemma: either store empty space or create an index to keep track of the nonempty cells or search for an alternative solution

# Data Quality

- Real world database are highly unprotected from noise, missing and inconsistent data due to their typically huge size and their possible origin from multiple, heterogeneous sources.
- Low quality data will lead to low quality mining results.
- Data pre-processing is required to handle these above mentioned facts.
- The methods for **data preprocessing** are organized into
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction
  - Data Discritization

# Data Cleaning

- Mostly concern with
  - Fill-in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Eliminate duplicate data
  - **Missing Data**
    - Data is not always available because many tuples may not have recorded values for several attributes such as age, income.
    - Missing data may be due to:
    - 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.
    - No change in recorded data.

# How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing. Not effective when the percentage of missing values per attribute varies considerably.
- Fill-in missing values manually: Tedious and infeasible task.
- Use a global constant to fill-in missing values.
- Use an attribute mean fill-in missing values belonging to the same class.
- Use the most probable value to fill-in missing value.

# Noisy Data

- Noisy data is a form of error because of random error in a measured variable.
- Incorrect attribute values may be due to:
- Faulty data collection instruments
- Data entry problem
- Data transmission problem
- Technology limitation
- Inconsistency in naming convention

**How to Handle Noisy Data**

- Clustering: Detect and remove outliers
- Regression: Smooth by fitting the data into regressi9on function
- Binning Method: First sort the data and partition into different boundaries with mean, median values.
- Combined computer and human inspection, doing so suspicious values are detected by human

# Outliers

- Outliers are a set of data points that are considerably dissimilar or inconsistent with the remaining data.
- In most of the cases they are inference of noise while in some cases they may actually carry valuable information.

Outliers can occur because of:

- Transient malfunction of data measurement
- Error in data transmission or transcription
- Changes in system behavior
- Data contamination from outside the population examined.
- Flaw in assumed theory

# Outliers

## How to Handle Outliers

- There are three fundamental approaches to the problem of outlier’s detection

- Type 1:
  - Determine the outliers with no prior knowledge of data. This is a learning approach analogous to unsupervised learning.
- Type 2:
  - Model with normality and abnormality. Analogous to supervised learning.
- Type 3:
  - Model with normality. Semi- supervised learning approach

# Data Integration

- Combines data from multiple sources into a coherent store.
- Integrate meta data from different sources (Schema Integration)
  - Problem: - Entity Identification Problem.
  - Different sources have different values for same attributes.
  - Data Redundancy
- These problems are mainly because of different representation, different scales etc.

**How to handle redundant data in data integration?**

- Redundant data may be able to be detected by correlation analysis.
- Step-wise and careful integration of data from multiple sources may help to improve mining speed and quality.

# Data Transformation

- Changing data from one form to another form.
- Approaches:
  - Smoothing: Remove noise from data.
  - Aggregation: Summarizations of data
  - Generalization: Hierarchy climbing of data
  - Normalization: Scaled to fall within a small specified range.

**Types**

- Min-Max Normalization:
  - $V' = ((V-min)/(max-min) * (new\_max - new\_min)) + new\_min$
- Z-Score Normalization:
  - $V' = (V-min)/ stand\_dev.$
- Normalization by decimal scaling:
  - $V' = V/ 10^j$ where j is the smallest integer such that max $(|V'|) < 1$

# Data Aggregation:

Combining two or more attributes (or objects) into a single attribute (or object).

- **Purpose**
  - Data reduction:  Reduce the number of attributes or objects
  - Change of scale: Cities aggregated into regions, states, countries, etc
  - More “stable” data:  Aggregated data tends to have less variability

# Data Reduction:

- Warehouse may store terabytes of data hence complex data mining may take a very long time to run on complete data set.

- Data reduction is the process of obtaining a reduced representation of data set that is much smaller in volume but yet produces the same or almost same analytical results.

- Different methods such as data sampling, dimensionality reduction, data cube, aggregation, discritization and hierarchy are used for data reduction.

- Data compression can also be used mostly in media files or data.

# Data Sampling:

- It is one of main method for data selection
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
- Often used for both preliminary investigation of data and the final data analysis.
- Important since obtaining entire set of data of interest is too expensive or time consuming.
- Sampling should be representative since it must represent approximately the same property as the original set of data.

# Sampling types

- ***Simple Random Sampling***: Equal probability of selecting any particular item.

- ***Sampling without replacement***: As each item is selected, it is removed from population.

- ***Sampling with replaceme***nt: Objects are not removed from the population as they are selected from the sample. The same objects can be picked-up more than once.

- ***Stratified Sampling***: Split the data into several partitions, then draw random samples from each partition.

# Data Discretization

- Convert continuous data into discrete data.
- Partition data into different classes.

**Two approaches are:**

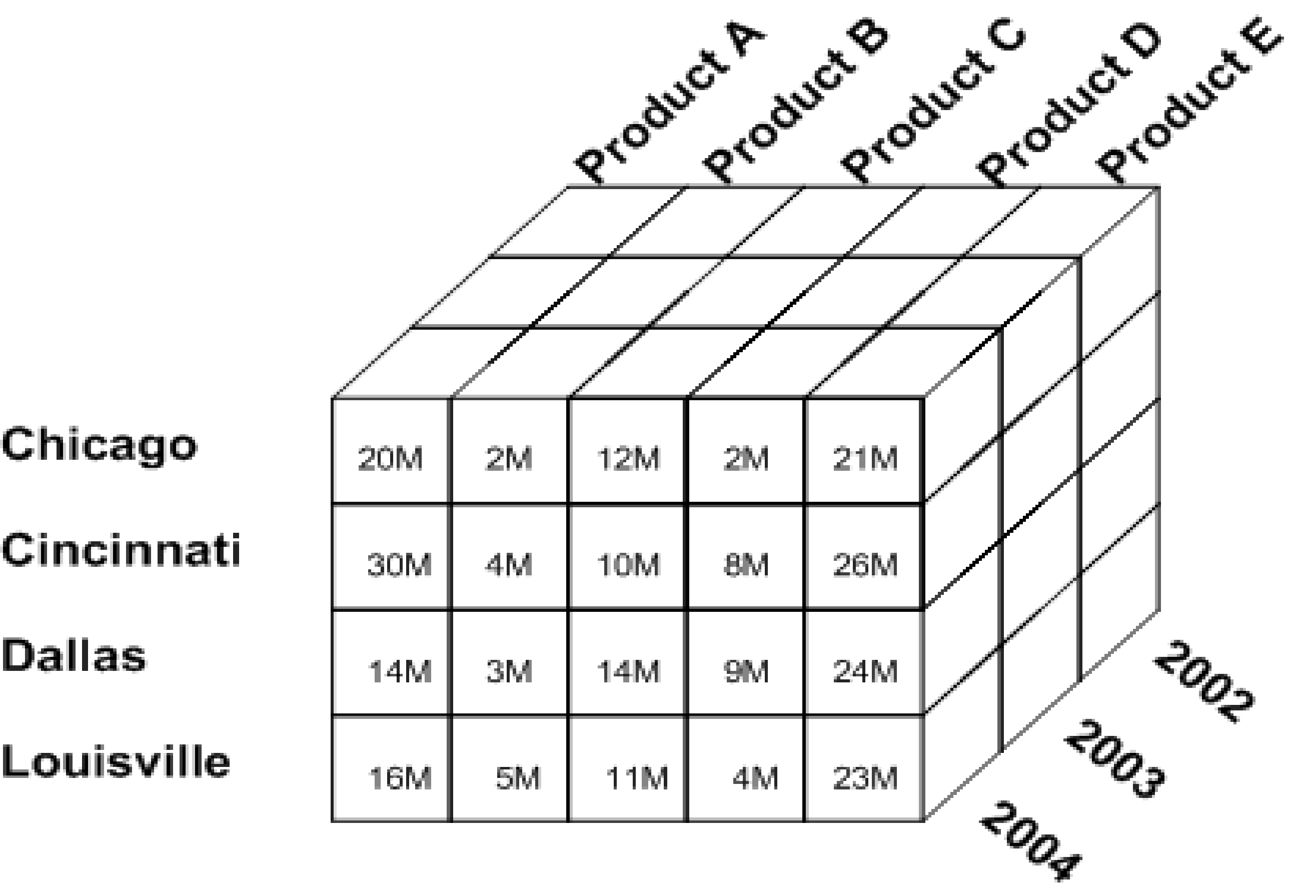
- **Equal width (distance) partitioning:**
  - It divides the range into N intervals of equal size.
  - If A and B are the lowest and the highest values of the attribute, the width of interval will be

    W = (A – B)/N.
  - The most straight forward approach for data discretization.
- **Equal depth (frequency) partitioning:**
  - It divides the range into N intervals, each containing approximately same number of samples.
  - Good data scaling
  - Managing categorical attributes can be tricky.

# OLAP

- OLAP stands for On-Line Analytical Processing.
- An OLAP cube is a data structure that allows fast analysis of data.
- OLAP tools were developed to solve multi-dimensional data analysis which stores their data in a special multi-dimensional format (data cube) with no updating facility.
- An OLAP toll doesn’t learn, it creates no new knowledge and they can’t reach new solutions.
- Information of multi-dimension nature can’t be easily analyzed when the table has the standard 2-D representation.
- A table with n- independent attributes can be seen as an n-dimensional space.
- It is required to explore the relationships between several dimensions and standard relational databases are not very good for this.

# OLAP Tool



| | Product A | Product B | Product C | Product D | Product E |
|---|---|---|---|---|---|
| Chicago | 20M | 2M | 12M | 2M | 21M |
| Cincinnati | 30M | 4M | 10M | 8M | 26M |
| Dallas | 14M | 3M | 14M | 9M | 24M |
| Louisville | 16M | 5M | 11M | 4M | 23M |

# OLAP Operations

- **Slicing:** A slice is a subset of multi-dimensional array corresponding to a single value for one or more members of the dimensions. Eg: Product A sales.
- **Dicing:** Dicing operation is the slice on more than two dimensions of data cube. (More than two consecutive slice). Eg: Product A sales in 2004.
- **Drill-Down:** Drill-down is specific analytical technique where the user navigates among levels of data ranging from the most summarized to the most detailed i.e. it navigates from less detailed data to more detailed data. Eg: Product A sales in Chicago in 2004.
- **Roll-Up:** Computing of all the data relationship for more than one or more dimensions i.e. summarization of data to one o more dimensions. Eg: Total Product.
- **Pivoting:** Pivoting is also called rotate operation. It rotates the data in order to provide an alternative presentation of data.

# OLTP (Online Transaction Processing)

- Used to carry out day to day business functions such as ERP (Enterprise Resource Planning), CRM ( Customer Relationship Planning)
- OLTP system solved a critical business problem of automating daily business functions and running real time report and analysis.

# OLAP Vs OLTP

| Facts | OLTP | OLAP |
|---|---|---|
| Source of Data | Operational Data | Data warehouse (From various database) |
| Purpose of data | Control and run fundamental business tasks | For planning, problem solving and decision support |
| Queries | Simple queries | Complex queries and algorithms |
| Processing Speed | Typically very fast | Depends on data size, techniques and algorithms |
| Space requirements | Can be relatively small | Larger due to aggregated databases |
| Database Design | Highly Normalized with many tables. | Typically denormalized with fewer tables. Use of star or snowflake schema. |

# Similarity and Dissimilarity of OLAP and OLTP

- **Similarity**
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- **Dissimilarity**
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies