Unit 2: Data Preprocessing

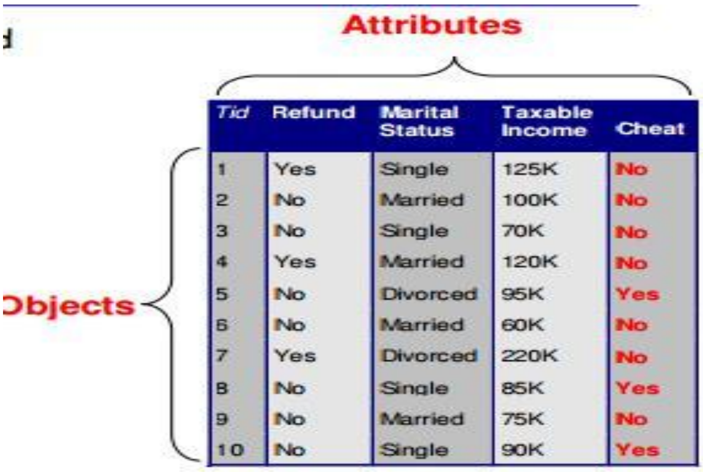
What is Data?

- Collection of data objects and their attributes.

• An attribute is a property or characteristic of an object

– Examples: eye color of a person, temperature, etc.

• A collection of attributes describe an object – Object is also known as record, point, case, sample, entity, or instance.


Attributes
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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

• There are different types of attributes

– **Nominal** : Examples: ID numbers, eye color, zip codes

–**Ordinal** :Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

– **Interval** : Examples: calendar dates, temperatures in Celsius or Fahrenheit.

– **Ratio** : In a ratio scale, numbers can be compared as multiples of one another. Thus one person can be twice as tall as another person. Thus the difference between a person of 35 and a person 38 is the same as the difference between people who are 12 and 15. Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

• The type of an attribute depends on which of the following properties it possesses:

-- **Distinctness**: = ≠

– **Order**: < >

– **Addition**: + -

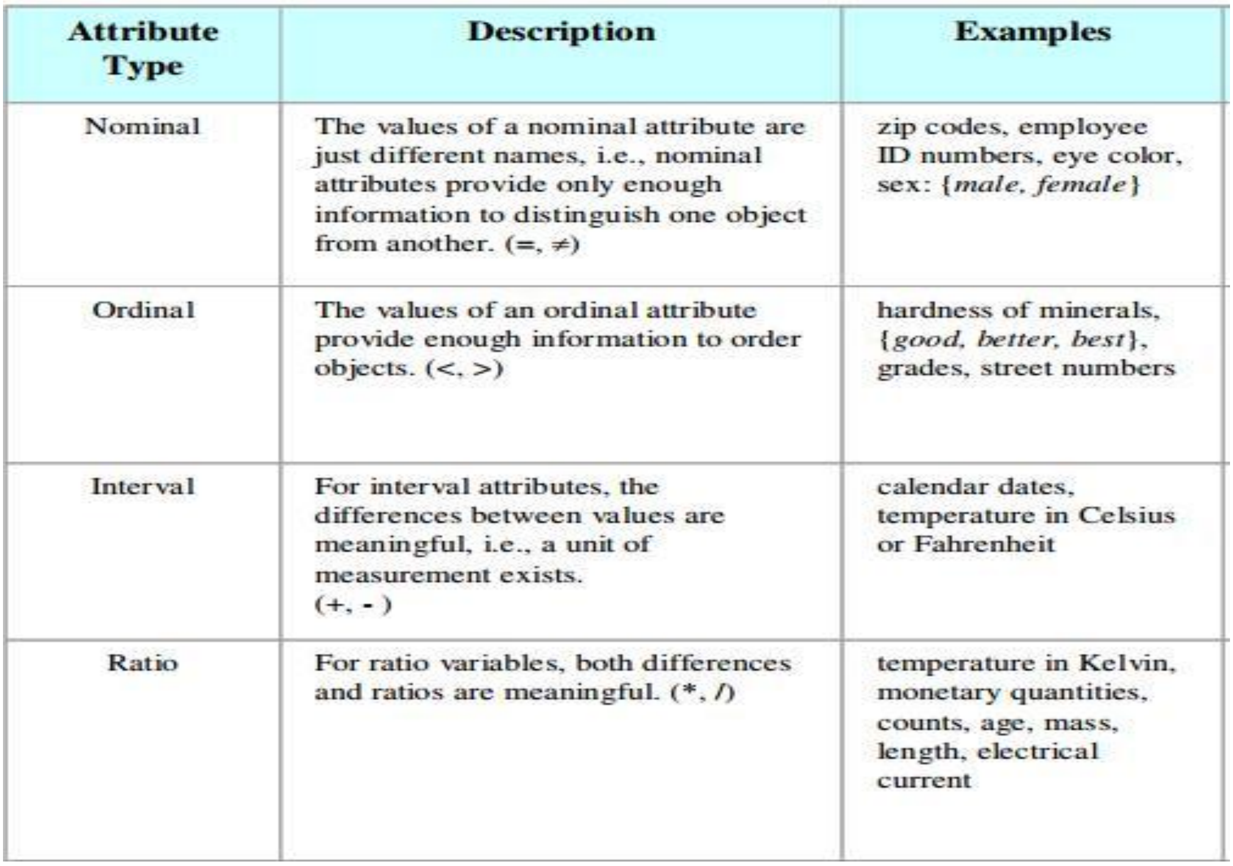
– **Multiplication**: * /

– Nominal attribute: distinctness

– Ordinal attribute: distinctness & order

– Interval attribute: distinctness, order & addition

– Ratio attribute: all 4 properties



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, sex: {*male, female*}
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. (+, -)	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

DISCRETE AND CONTINUOUS ATTRIBUTES

- **Discrete Attribute** [Nominal and Ordinal]

–Has only a finite 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** [interval and ratio]

–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.

Data Pre-processing

-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. Data in **the real world is dirty**:

- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=" "
- **noisy**: containing errors or outliers
  - e.g., Salary="-10"
- **inconsistent**: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"

- Low quality data will lead to low quality mining results.

- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.

-Data warehouse needs consistent integration of quality data

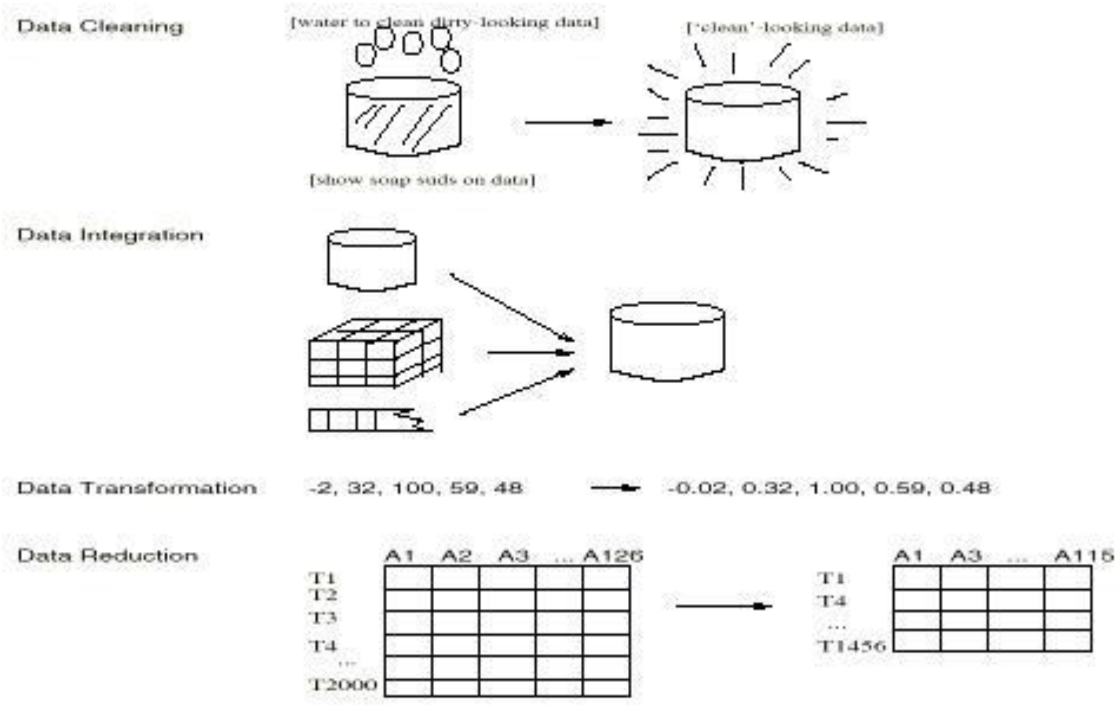
- Data pre-processing is required to handle these above mentioned facts.

The methods for data preprocessing are organized into (**Major Tasks in Data Preprocessing**)

1. **Data cleaning**

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

2. **Data integration**
   - Integration of multiple databases, data cubes, or files
3. **Data transformation**
   - Normalization and aggregation
4. **Data reduction**
   - Obtains reduced representation in volume but produces the same or similar analytical results
5. **Data discretization**
   - Part of data reduction but with particular importance, especially for numerical data



Data Cleaning

Mostly concern with

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

     Data is not always available

E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

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

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. Eg: unknown
- Use an attribute mean fill-in missing values belonging to the same class.
- Use the most probable value to fill-in missing value: inference-based such as Bayesian formula or decision tree.

b. Noisy Data

- Noisy data is a form of error because of random error in a measured variable.
- Incorrect attribute values may be due to:

- Clustering: Detect and remove outliers
- Regression: Smooth by fitting the data into regression 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 (e.g., deal with possible outliers)

Binning Methods for Data Smoothing:

❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

c. 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.

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:

i. Smoothing: Remove noise from data.

ii. Aggregation: Summarizations of data

iii. Generalization: Hierarchy climbing of data -> low-level =>high level concepts e.g. age => youth, middle-aged, senior

iv. Normalization: ->attribute data are scaled into specified range such as -1.0 to 1.0 or 0.0 to 1.0 (e. g. how??)

E.g.

$$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$$

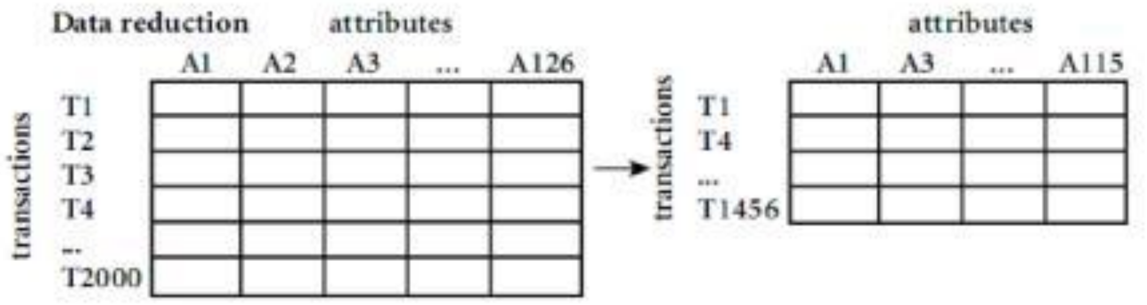
Data Aggregation:

- 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

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



Data Discretization:

- Convert continuous data into discrete data.

OLAP Tool

Definition: OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view. For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September, and then see a comparison of other product sales in Florida in the same time period. To facilitate this kind of analysis, OLAP data is stored in a multidimensional database.

- 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.
- 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.