جامعة الطائف
TAIF UNIVERSITY

DATA MINING

Assoc. Prof. Dr. Salha Alzahrani

College of Computers and Information Technology
Taif University
Saudi Arabia

s.zahrani@tu.edu.sa

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Recap of Lecture 1

- Black-Box design of data mining
- Motivation: Why data mining?
- Evolution of sciences
- Evolution of database technology
- What is data mining?
- Knowledge discovery, Knowledge discovery in databases
- Data mining and business intelligence
- Why not traditional data analysis?
- Multi-dimensional view of data mining
- Conferences and Journals on data mining
- Where to Find References? DBLP, CiteSeer, Google
- Applications of data mining

Data Processing | Data Cleaning, Preparation, Dealing with Missing Data, and Attributes Reduction

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Content of Lecture 2

- Standard Formulation
- Types of Variables
- Categorical and Continuous Attributes
- Data Preparation
  - Data Cleaning
  - Missing Values
  - Attributes Reduction
- Summary & Checklist.

Data Processing | Standard formulation for data

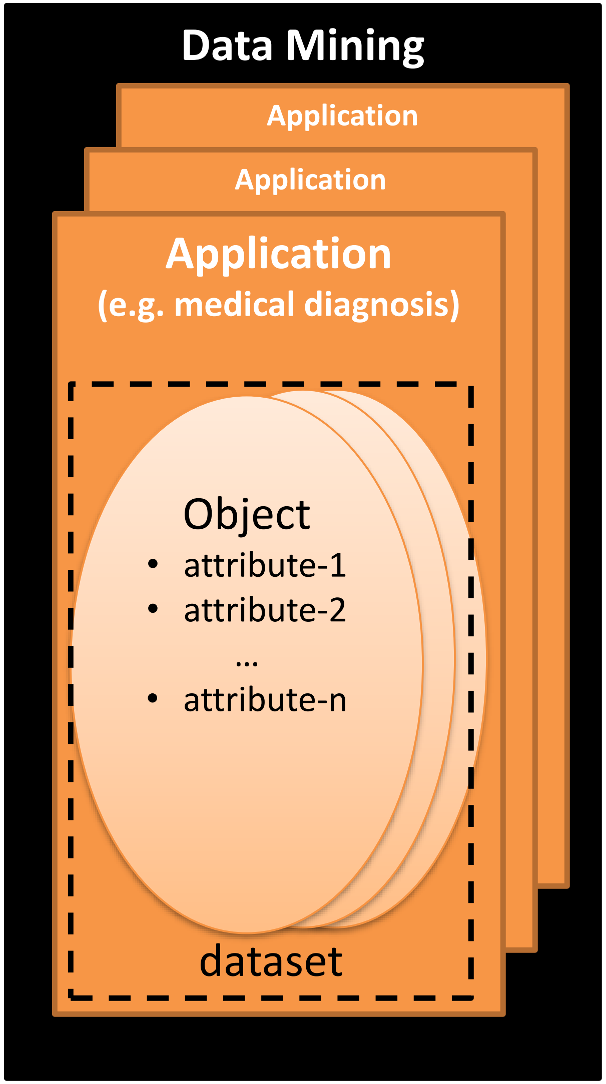
- For any data mining application, we have a ***universe of objects*** that are of interest.

- The universe of objects is normally very large and we have only a small part of it.

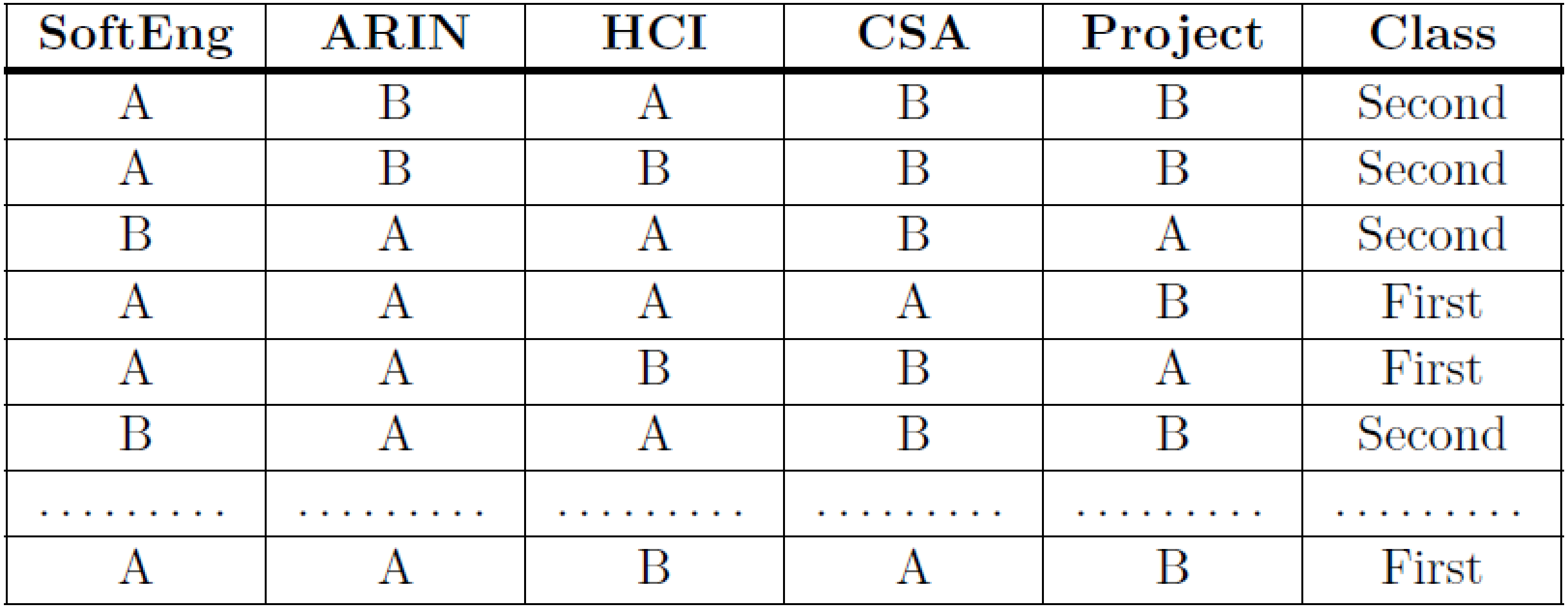
- Each object is described by a number of ***variables*** that correspond to its properties. In data mining, variables are often called ***attributes***.

- The set of variable values corresponding to each of the objects is called a ***record*** or (more commonly) an ***instance***.

- The complete set of data available to us for an application is called a ***dataset***.



Data Processing | Standard Formulation: Dataset Example



SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
....	....	....	....	....	....
A	A	B	A	B	First

Figure 1.1 The Degrees Dataset

Data Processing | Labelled and unlabelled data

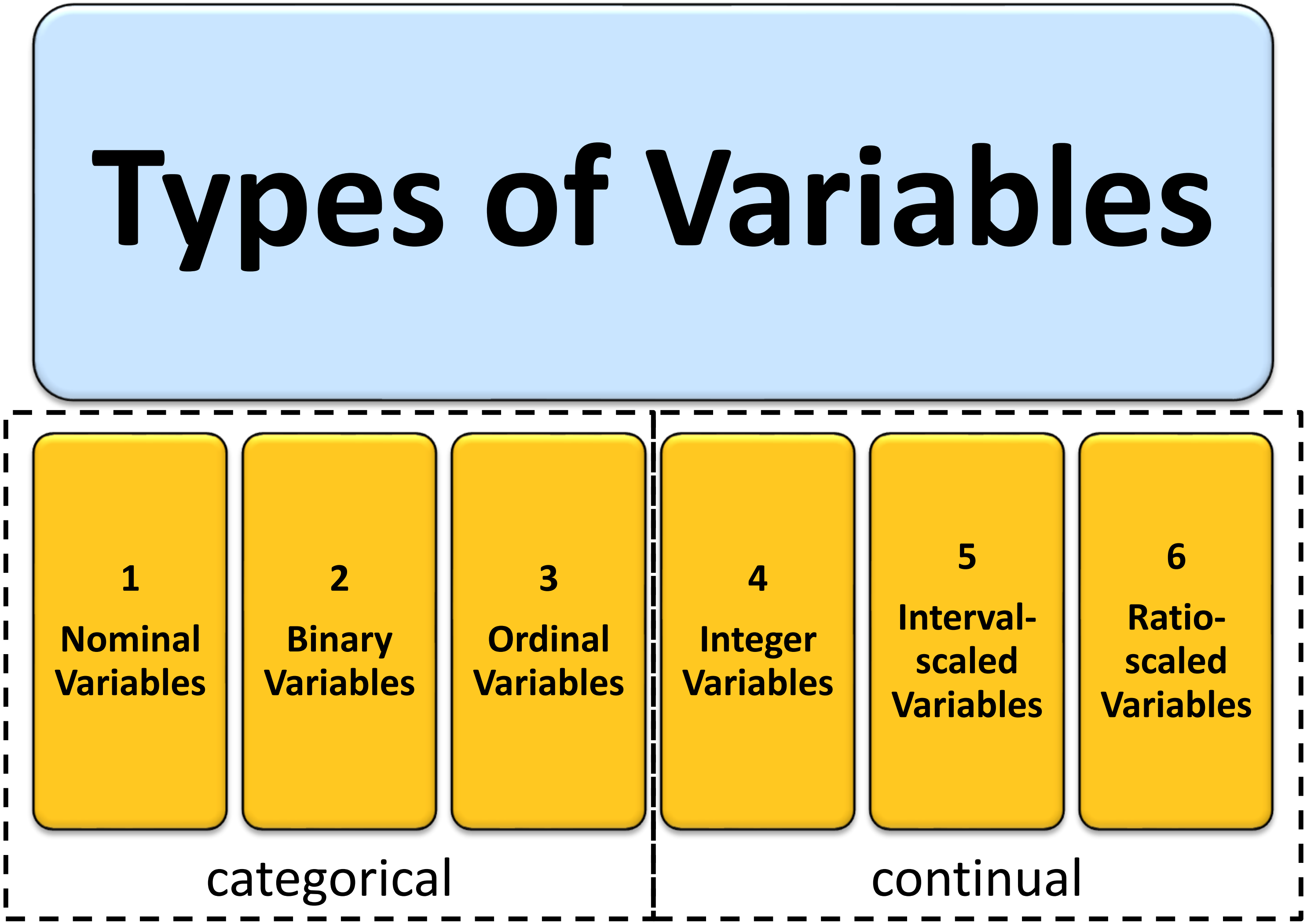
- **Labelled data** is the dataset where one attribute is given special significance and the aim is to predict its value.

- This attribute has a standard name **'class'**. (e.g. Degrees dataset).

- **Unlabelled data** is the dataset that do not contain any significant attribute.

- For many applications it is helpful to have a third category of attribute, the **'ignore'** attribute, corresponding to variables that are of no significance for the application, for example the name of a patient in a hospital or the serial number of an instance, but which we do not wish to delete from the dataset.

Data Processing | Types of Variables



Types of Variables

categorical

1 Nominal Variables

2 Binary Variables

3 Ordinal Variables

continual

4 Integer Variables

5 Interval-scaled Variables

6 Ratio-scaled Variables

Data Processing | Nominal Variables

1. Nominal Variables

- A nominal variable is used to put objects into categories, e.g. the name or colour of an object.

- A nominal variable may be numerical in form, but the numerical values have no mathematical interpretation.

- For example we might label 10 people as numbers 1*,* 2*,* 3*, . . . ,* 10, but any arithmetic with such values, e.g. 1 + 2 = 3 would be meaningless.

Data Processing | Binary Variables

2. Binary Variables

- A binary variable is a special case of a nominal variable that takes only two possible values: true or false, 1 or 0.

Data Processing | Ordinal Variables

3. Ordinal Variables

- Ordinal variables are similar to nominal variables, except that an ordinal variable has values that can be arranged in a meaningful order, e.g. small, medium, large.

4. Integer Variables

- Integer variables are ones that take values that are genuine integers, for example ‘number of children’.
- Unlike nominal variables that are numerical in form, arithmetic with integer variables is meaningful (1 child + 2 children = 3 children etc.).

Data Processing | Interval-scaled Variables

5. Interval-scaled Variables

- Interval-scaled variables are variables that take numerical values which are measured at equal intervals from a zero point or origin. However, the origin does not imply a true absence of the measured characteristic.

- Two well-known examples of interval-scaled variables are the Fahrenheit and Celsius temperature scales.

- To say that one temperature measured in degrees Celsius is greater than another or greater than a constant value such as 25 is clearly meaningful.

Data Processing | Ratio-scaled Variables

6. Ratio-scaled Variables

- Ratio-scaled variables are similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic.
- For example, molecular weight, price, etc.
- A weight of 10 kg is twice one of 5 kg, a price of 100 dollars is twice a price of 50 dollars etc.

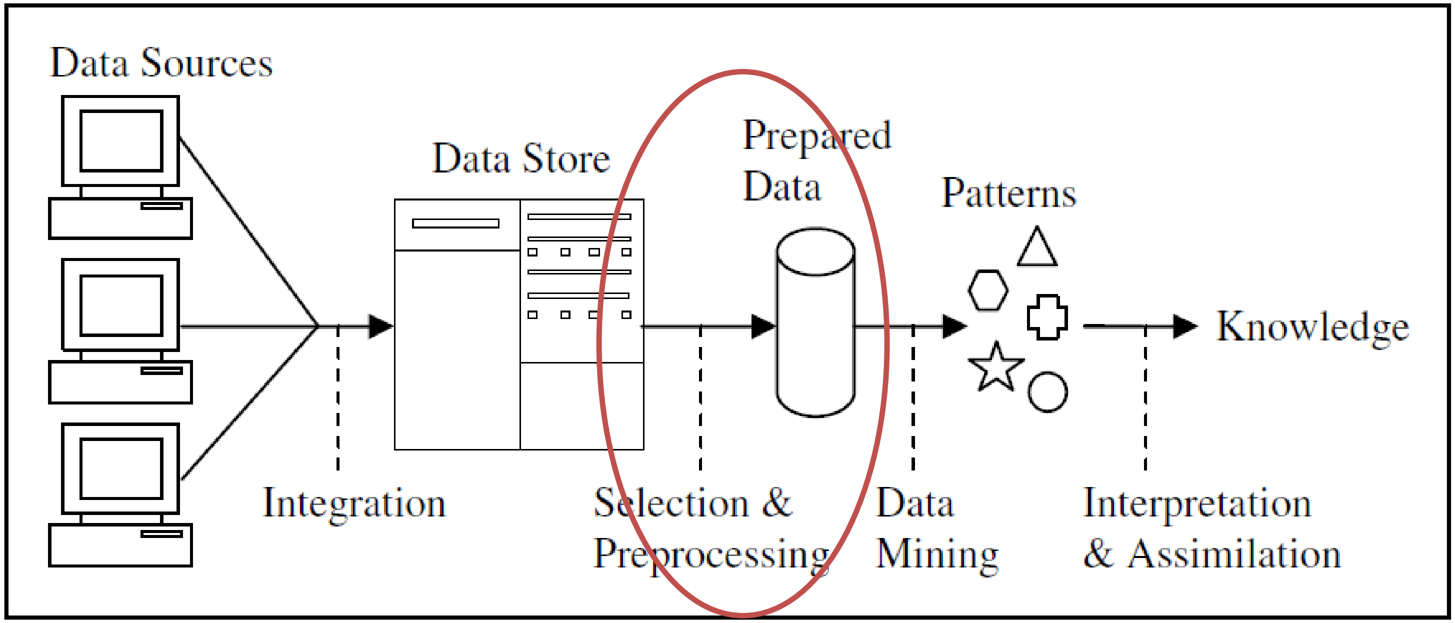
Categorical and Continuous Attributes

Many practical data mining systems divide attributes into two types:
– **categorical** corresponding to nominal, binary and ordinal variables.
– **continuous** corresponding to integer, interval-scaled and ratio-scaled variables.

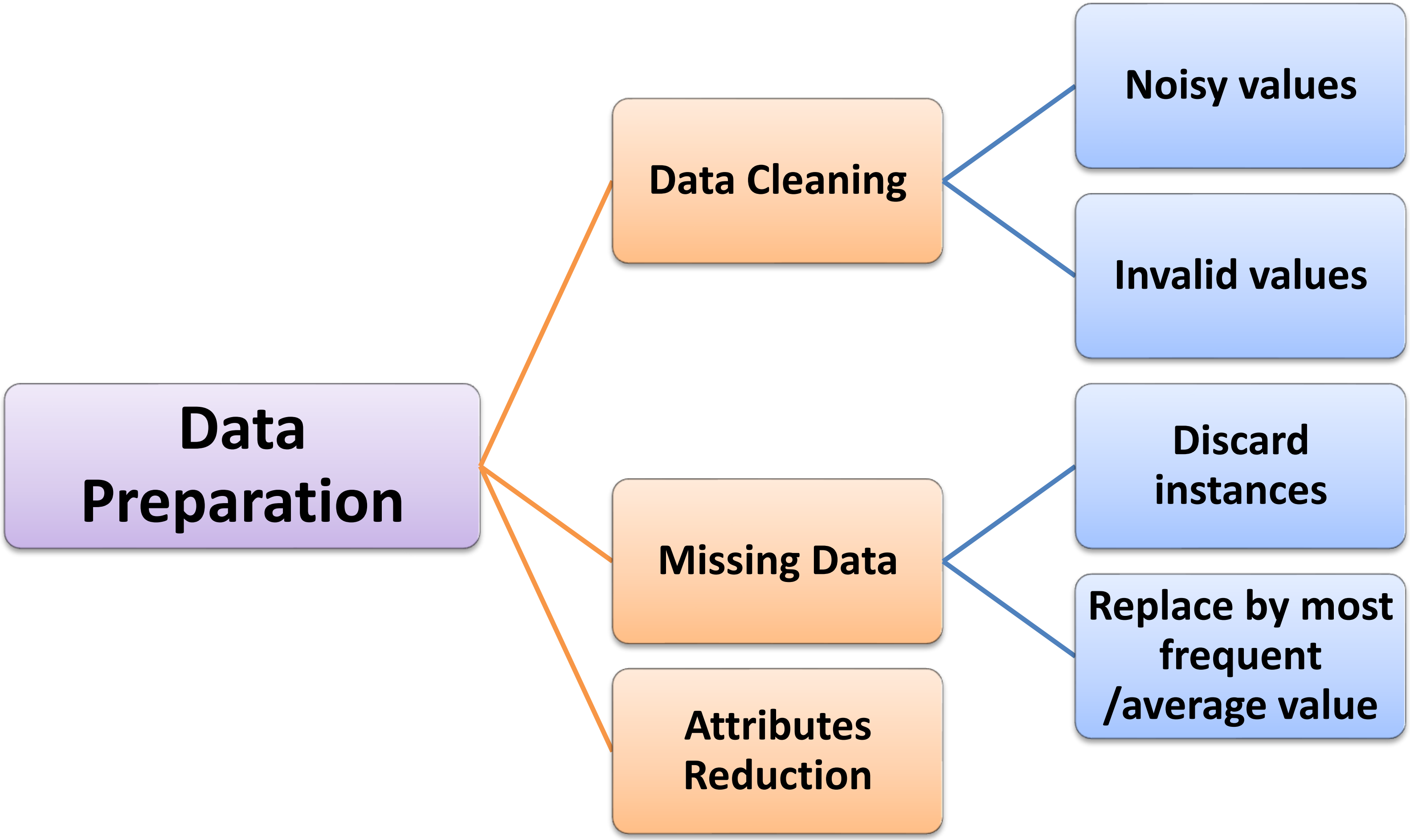
- It is important to choose methods that are appropriate to the types of variable stored for a particular application.

Data Preparation

A step before data mining: it is important to get the data into a standard form in which it can be analysed.



Data Processing | Data Preparation



Data Preparation

Data Cleaning

Noisy values

Invalid values

Missing Data

Discard instances

Replace by most frequent /average value

Attributes Reduction

Data Processing | Data Preparation >>Data Cleaning

Data Cleaning

- Even when the data is in the standard form it cannot be assumed that it is error free.

- **How errors occur?**

  In real-world datasets, erroneous values can be recorded for a variety of reasons, including measurement errors, subjective judgements and malfunctioning/misuse of automatic recording equipment.

Data Processing | Data Preparation >>Data Cleaning

What types of errors?

Erroneous values can be divided into those which are possible values of the attribute and those which are not.

- A ***noisy* value** is the value that is valid for the dataset, but is incorrectly recorded. For example, the number 69.72 may accidentally be entered as 6.972, or a categorical attribute value such as *brown* may accidentally be recorded as another of the possible values, such as *blue*.
- An ***invalid* value** is the value that is invalid for the dataset, such as 69.7X for 6.972 or *bbrown* for *brown*.

Data Processing | Data Preparation >>Data Cleaning

How to clean data?

- Using software tools, especially to give an overall visual impression of the data, when some anomalous values or unexpected concentrations of values may stand out.

- Using some basic analysis of the values. For example, sorting the values into ascending order (which for fairly small datasets can be accomplished using just a standard spreadsheet) may reveal unexpected results.

Missing Values

- In many real-world datasets, data values are not recorded for all attributes.
- **How missing data occur?**
  - This can happen simply because there are some attributes that are not applicable for some instances (e.g. certain medical data may only be meaningful for female patients).
  - It can also happen that there are attribute values that should be recorded that are missing. This can occur for several reasons, for example, a malfunction of the equipment used to record the data.

Data Processing | Missing Values >> Discard instances

How to deal with missing data?

Strategy 1: Discard Instances

- This is the simplest strategy: delete all instances where there is at least one missing value and use the remainder.

- **Advantage:**

  - This strategy is a very **conservative** one, which has the advantage of avoiding introducing any data errors.

- **Disadvantage:**

  - Discarding instances may **damage the reliability** of the results derived from the data.

  - **Not usable** when all or a high proportion of all the instances have missing values.

Data Processing | Missing Values >> Replace by Most Frequent/Average Value

How to deal with missing data?

Strategy 2: Replace by Most Frequent/Average Value

- Another approach is to estimate each of the missing values using the values that are present in the dataset.
  - For **categorical attribute**: use its most frequently occurring value.
  - For **continuous attribute**: use the average value.

- **Advantage:**

  - Replacing missing values by most frequent/average value may **preserve the reliability** of the results derived from the data.

- **Disadvantage:**

  - This strategy may **introduce noise** into the data if the proportion of missing values for a variable is big.

How big number of attributes may occur?

- In some data mining applications, the availability of larger storage capacity has led to large numbers of attribute values being stored for every instance, e.g. information about all the purchases made by a supermarket customer for three months.

- For some datasets, there can be substantially more attributes than there are instances. Many of those attributes are **irrelevant** to the data mining application.

Data Processing | Data Preparation >> Attributes Reduction

Why we need to reduce the number of attributes?

- Irrelevant attributes will place an unnecessary computational overhead on any data mining algorithm.
- At worst, they may cause the algorithm to give poor results.

How we can reduce the number of attributes?

- There are several methods in which the number of attributes (or 'features') can be reduced before a dataset is processed. These methods are called ***feature reduction*** or ***dimension reduction*** such as principle component analysis (PCA) and others.

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Summary & Checklist

- ☑ Standard Formulation
- ☑ Types of Variables
- ☑ Categorical and Continuous Attributes
- ☑ Data Preparation
  - ☑ Data Cleaning
  - ☑ Missing Values
  - ☑ Attributes Reduction

Reminder | Next Lecture !

Next Lecture…

Classification using Naïve Bayes Algorithm (Ch. 2)

- ***Be ready!***
- ***Download & print the lecture notes before your class.***

Thank You !

✉ s.zahrani@tu.edu.sa