Introduction to Data Mining

Course Code: CSC 4139 Course Title: Data Warehouse and Data Mining

Dept. of Computer Science
Faculty of Science and Technology

Lecturer No:	**1 & 2**	**Week No:**	**1**	**Semester:**	**Summer20-21**
Lecturer:	*Dr. Md Mahbub Chowdhury Mishu*				

Introduction to Data Mining

- Lecture Objectives and Outcomes:
  - Introduce the “big picture” of data mining
  - Explore the data mining process, and
  - Various applications of data mining

Why Data Mining

- The Data Explosion:
  - Modern computer systems are accumulating data at an almost unimaginable rate and from a very wide variety of sources: from point-of-sale machines in the high street to machines logging every cheque clearance, bank cash withdrawal and credit card transaction, to Earth observation satellites in space.
  - Some examples will serve to give an indication of the volumes of data involved:

Why Data Mining

- The current NASA Earth observation satellites generate a terabyte (i.e. 109 bytes) of data every day. This is more than the total amount of data ever transmitted by all previous observation satellites.
- The Human Genome project is storing thousands of bytes for each of several billion genetic bases.
- As long ago as 1990, the US Census collected over a million million bytes of data.
- Many companies maintain large Data Warehouses of customer transactions. A fairly small data warehouse might contain more than a hundred million transactions.

Applications of Data Mining

- There is a rapidly growing body of successful applications in a wide range of areas as diverse as:
  - analysis of organic compounds
  - automatic abstracting
  - credit card fraud detection
  - electric load prediction
  - financial forecasting
  - medical diagnosis
  - predicting share of television audiences
  - product design
  - real estate valuation
  - targeted marketing
  - thermal power plant optimisation

Applications of Data Mining

- toxic hazard analysis
- weather forecasting
- a supermarket chain mines its customer transactions data to optimise targeting of high value customers
- a credit card company can use its data warehouse of customer transactions for fraud detection
- a major hotel chain can use survey databases to identify attributes of a ‘high-value’ prospect
- predicting the probability of default for consumer loan applications by improving the ability to predict bad loans
- reducing fabrication flaws in VLSI chips

Applications of Data Mining

- data mining systems can sift through vast quantities of data collected during the semiconductor fabrication process to identify conditions that are causing yield problems
- predicting audience share for television programmes, allowing television executives to arrange show schedules to maximise market share and increase advertising revenues
- predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care.

The world is becoming '*data rich but knowledge poor*'.

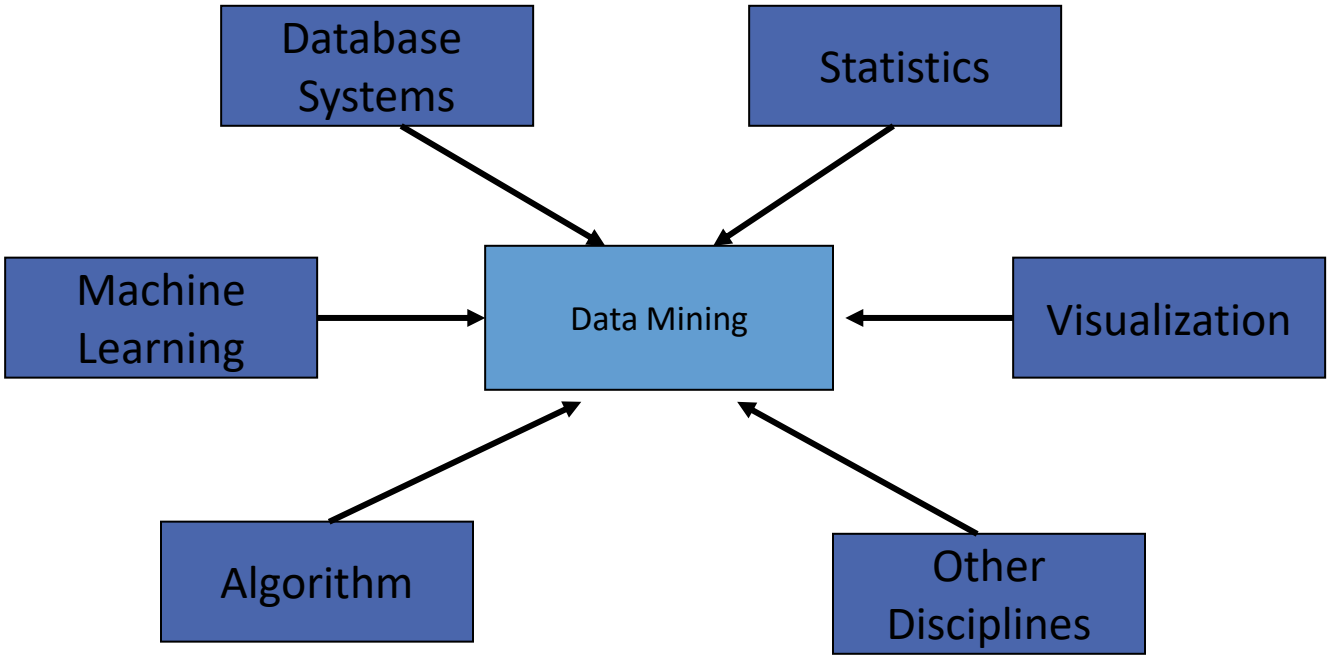
What is Data Mining

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative name
  - Knowledge discovery in databases (KDD)
- Is everything “data mining”?

What is NOT Data Mining

- Look up phone number in phone directory
- Query a Web search engine for information about "Amazon"

Data Mining: Union of Multiple Disciplines



Applications of Data Mining

Applications can be divided into four main types:

- Classification
- Numeric prediction
- Association
- Clustering

Before a brief explanation of these types we need to distinguish between two types of data

- Labelled and
- Unlabelled Data

Labelled and Unlabelled Data

- Consider a dataset of examples (called *instances*)
  - Comprises **variables** or ***attributes***

- There are two types of data, which are treated in fundamentally

different ways.

A. First type there is a **specially designated attribute** – the aim is to use the data given to predict the value of ***that*** attribute for instances that have not yet been seen. Data of this kind is called **labelled**.
B. Data that does not have any specially designated attribute is called **unlabelled**

Labelled and Unlabeled Data

- **Labelled** Data
  - Data mining using labelled data is known as **supervised** learning.
  - If the designated attribute is **categorical**, i.e. it must take one of a number of distinct values such as ‘very good’, ‘good’ or ‘poor’, or (in an object recognition application) ‘car’, ‘bicycle’, ‘person’, ‘bus’ or ‘taxi’ the task is called **classification**.
  - If the designated attribute is **numerical**, e.g. the expected sale price of a house or the opening price of a share on tomorrow’s stock market, the task is called **regression**.

Labelled and Unlabeled Data

- **Unlabeled** Data
  - Data mining of unlabeled data is known as **unsupervised** learning.
  - Here the aim is simply to extract the most information we can from the data

Supervised Learning: Classification

- Classification is one of the most common mining applications
- Example: Dataset containing students' grades on five subjects and their overall degree classifications
- Predicting the **classification**, given only their grade 'profiles'.
  - Nearest Neighbor Matching
  - Classification Rules
  - Classification Tree or Decision Tree

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

Unsupervised Learning: Association Rules

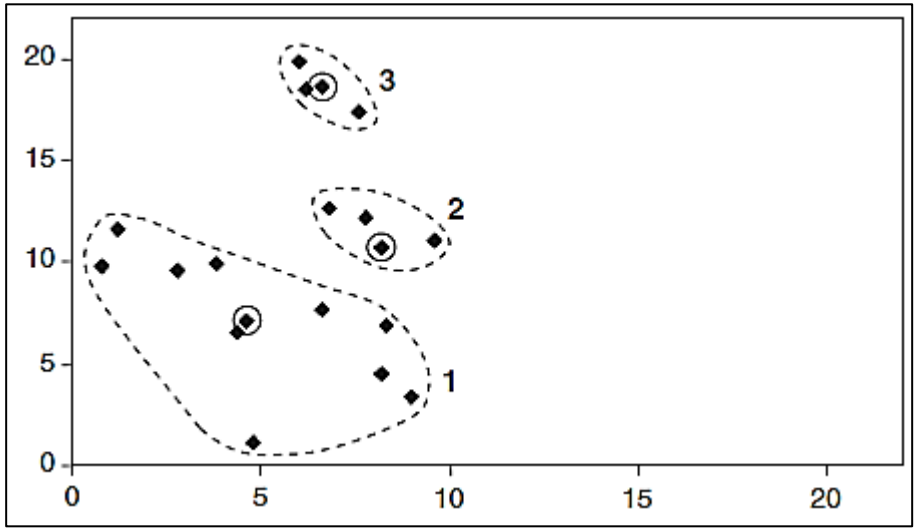
- Find relationship that exists amongst the values of variables, in the form of rules known as association rules.
- A common form of this type of application is called ‘market basket analysis’.

IF variable_1 $>$ 85 and switch_6 = open
THEN variable_23 $<$ 47.5 and switch_8 = closed (probability = 0.8)

IF cheese AND milk THEN bread (probability = 0.7)

Unsupervised Learning: Clustering

- Clustering algorithms examine data to find groups of items that are similar.
- For example, an insurance company might group customers according to income, age, types of policy purchased or prior claims experience.



Data for Data Mining

- Lecture Objectives and Outcomes:
  - The standard formulation for the data input to data mining algorithms
  - Distinguish between different types of variables and to consider issues relating to the preparation of data prior to use, particularly the presence of missing data values and noise
  - The UCI repository and Kaggle datasets are introduced

Getting to Know Your Data

- It's tempting to jump straight into mining, but first, we need to get the data ready. This involves having a closer look at attributes and data values. Real-world data are typically noisy, enormous in volume (often several gigabytes or more) and may originate from a hodgepodge of heterogenous sources.
- Knowledge about your data is useful for data preprocessing, the first major task of the data mining process. You will want to know the following:
- What are the types of attributes or fields that make up your data?
- What kind of values does each attribute have?
- Which attributes are discrete, and which are continuous-valued?

Getting to Know Your Data

- What do the data look like?
- How are the values distributed?
- Are there ways we can visualize the data to get a better sense of it all?
- Can we spot any outliers?
- Can we measure the similarity of some data objects with respect to others?
- Gaining such insight into the data will help with the subsequent analysis.

Basic

- **Data**
  - *a "given"* or *a fact* that represents something in real world
  - raw materials, can be processed, structured or unstructured
  - Data are elements of analysis
- **Information**
  - Data that have *meaning in context*
  - Data related
  - Data after manipulation
- **Knowledge**
  - familiarity, awareness and understanding of someone or something
  - acquired through experience or learning
  - it is a concept mainly for humans unlike data and information.

- Knowledge is not information and information is not data.
- Knowledge is derived from information in the same way information is derived from data.

Key properties & Scope

- **Key Properties**
  - Automatic discovery of patterns (from given/store data)
  - Prediction of likely outcomes (from given/store data)
  - Creation of actionable information (from given/store data)
  - Focus on large data sets and databases

- **Scope**
  - Automated prediction of trends and behaviors
  - Automated discovery of previously unknown patterns

Basic Terminology

- **Data Objects**
- Data sets are made up of data objects. A data object represents an entity—in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a university database, the objects may be students, professors, and courses.
- Data objects are typically described by attributes.
- Data objects can also be referred to as samples, examples, instances, data points, or objects.
- If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

Basic Terminology

- **Attribute/Variable**
  - Each object is described by a number of variables that correspond to its properties. In data mining variables are often called attributes. Examples: eye color of a person
  - Categorical (qualitative):Categorical variables take on values that are names or labels. The color of a ball (e.g., red, green, blue)
  - Continuous (quantitative) : Quantitative variables are numerical. They represent a measurable quantity. when we speak of the population of a city, we are talking about the number of people in the city.

Basic Terminology

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

- **Dataset**
  - The complete set of data available to us for an application is called a dataset. A dataset is often depicted as a table, with each row representing an instance. Each column contains the value of one of the variables (attributes) for each of the instances.

- **Class attribute**

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

Types of Attribute/Variable

- Mainly six types of Attribute
  - **Nominal**
    - Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order.
    - A variable used to put objects into categories, e.g. the name or color of an object.
    - A nominal variable may be numerical in form, but the numerical values have no mathematical interpretation.

Types of Attribute/Variable

- **Example:**
  - Suppose that hair color and marital status are two attributes describing person objects.
  - In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white.
  - The attribute marital status can take on the values single, married, divorced, and widowed.
  - Both hair color and marital status are nominal attributes.
  - Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on.

Types of Attribute/Variable

- **Binary**
  - A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.
  - Binary attributes are referred to as Boolean if the two states correspond to true and false.
  - A binary variable is a special case of a nominal variable that takes only two possible values: true or false, 1 or 0 etc.

Types of Attribute/Variable

- **Example:**
  - Given the attribute smoker describing a patient object, 1 indicates that the patient smokes, while 0 indicates that the patient does not.
  - Similarly, suppose the patient undergoes a medical test that has two possible outcomes. The attribute medical test is binary, where a value of 1 means the result of the test for the patient is positive, while 0 means the result is negative.
  - A binary attribute is symmetric if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1. One such example could be the attribute gender having the states male and female.

Types of Attribute/Variable

- **Example:**
  - A binary attribute is asymmetric if the outcomes of the states are not equally important, such as the positive and negative outcomes of a medical test for HIV. By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).

Types of Attribute/Variable

- **Ordinal**
  - Ordinal variables are similar to nominal variables, except that an ordinal variable has values that can be arranged in a meaningful order/rank, e.g. small, medium, large. But the magnitude between successive values is not known.

- **Example:**
  - Suppose that drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large. The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values how much bigger, say, a medium is than a large.
  - Other examples of ordinal attributes include grade (e.g., A+, A, A-, B+, and so on) and professional rank. Professional ranks can be enumerated in a sequential order: for example, assistant, associate, and full for professor.

Types of Attribute/Variable

- **Integer/Numeric Attributes**
  - A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values.
  - Numeric attributes can be interval-scaled or ratio-scaled.

Types of Attribute/Variable

- **Interval-Scaled**
  - Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative.
  - Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.

Types of Attribute/Variable

- **Example:**
  - A temperature attribute is interval-scaled.
  - Suppose that we have the outdoor temperature value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to temperature. In addition, we can quantify the difference between values. For example, a temperature of 20° C is five degrees higher than a temperature of 15° C.
  - Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart. Temperatures in Celsius or Fahrenheit

Types of Attribute/Variable

- **Example:**
  - Temperatures in Celsius and Fahrenheit do not have a true zero-point, that is, neither 0 ° C nor 0 ° F indicates “no temperature.” (On the Celsius scale, for example, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure.) Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another. Without a true zero, we cannot say, for instance, that 10 ° C is twice as warm as 5 ° C. That is, we cannot speak of the values in terms of ratios.
  - Similarly, there is no true zero-point for calendar dates.

Types of Attribute/Variable

- **Ratio-Scaled**
  - A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
  - In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.
- **Example:**
  - Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K) temperature scale has what is considered a true zero-point 0 ° K = -273.15 ° C): It is the point at which the particles that comprise matter have zero kinetic energy.
  - Other examples of ratio-scaled attributes include count attributes such as years of experience (e.g., the objects are employees) and number of words (e.g., the objects are documents). Additional examples include attributes to measure weight, height, latitude and longitude coordinates, and monetary quantities.

Types of Attribute/Variable

- **Ignore attribute**
  - Corresponding to variables that are of no significance for the application

Types of Attribute/Variable

- **Categorical and Continuous Attributes:**
  - Although the distinction between different categories of variable can be important in some cases, many practical data mining systems divide attributes into just 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 Types

- Two types of data: Labelled Data & Unlabeled Data

- **Labelled data**
  - Specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen.
  - **SoftEng=A, ARIN=A, HCI=B, CSA=B, Project=A,**
    THEN Class = ?

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

Data Types

- **Unlabelled data**
  - Data that does not have any specially designated attribute is called unlabelled.
  - Here the aim is simply to extract the most information we can from the data available.

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

Date	Name	Shop Name	Products	…
25/09/2014	Marcos	ABC		…
27/09/2014	Marcos	ABC	Pen, Bread, Cake	…
05/10/2014	Marcos	ABC	Bread, Cheese	…

Learning Methods

- **Supervised Learning**
  - Data mining using labelled data is known as supervised learning.

- **Classification**
  - If the designated attribute is categorical, the task is called classification.
  - Classification is one form of prediction, where the value to be predicted is a label.
  - a hospital may want to classify medical patients into those who are at high, medium or low risk
  - we may wish to classify a student project as distinction, merit, pass or fail
  - Nearest Neighbour Matching, Classification Rules, Classification Tree, …

Learning Methods

- **Numerical Prediction (Regression)**
  - If the designated attribute is numerical, the task is called numerical prediction (regression).
  - Numerical prediction (often called regression) is another. In this case we wish to predict a numerical value, such as a company's profits or a share price.
  - A very popular way of doing this is to use a Neural Network

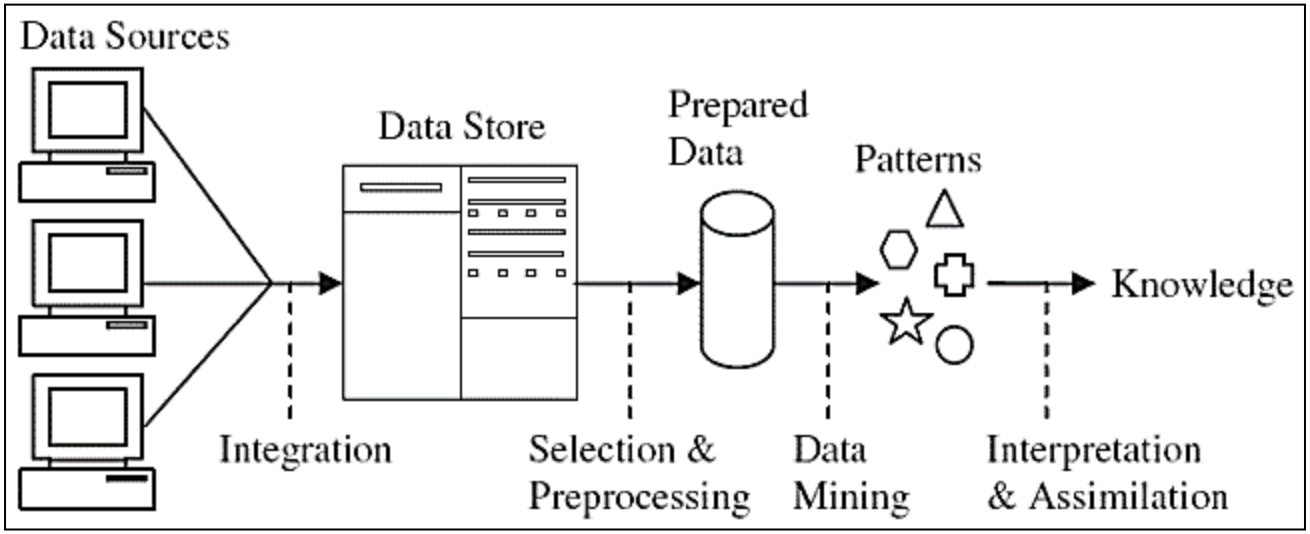
Learning Methods

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

Learning Methods

- **Unsupervised Learning**
  - Data mining using unlabelled data is known as unsupervised learning.
  - **Association Rules**
    - Sometimes we wish to use a training set to find any relationship that exists amongst the values of variables, generally in the form of rules known as association rules.
    - APRIORI
    - Market Basket Analysis
  - **Clustering**
    - Clustering algorithms examine data to find groups of items that are similar.
    - K-Means Clustering, Agglomerative Hierarchical Clustering

Knowledge Discovery Process



Data Preparation

- **Data Cleaning**
  - Even when the data is in the standard form it cannot be assumed that it is error free.
  - In real-world datasets erroneous values can be recorded for a variety of reasons, including measurement errors, subjective judgements and malfunctioning or misuse of automatic recording equipment.
  - **Noisy value**
    - a noisy value to mean one that is valid for the dataset, but is incorrectly recorded
    - 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.
  - **Invalid value**
    - 69.7X for 6.972 or bbrown for brown
    - An invalid value can easily be detected and either corrected or rejected

Data Preparation

- **Missing Values**
  - In many real-world datasets data values are not recorded for all attributes. This can happen simply because there are some attributes that are not applicable for some instances, a malfunction of the equipment used to record the data, a data collection form to which additional fields were added after some data had been collected, information that could not be obtained, e.g. about a hospital patient

- **Discard Instances**
  - This is the simplest strategy: delete all instances where there is at least one missing value and use the remainder.
  - It has the advantage of avoiding introducing any data errors. Its disadvantage is that discarding data may damage the reliability of the results derived from the data

Data Preparation

- **Replace by Most Frequent/Average Value**
  - A less cautious strategy is to estimate each of the missing values using the values that are present in the dataset.
  - A straightforward but effective way of doing this for a categorical attribute is to use its most frequently occurring (non-missing) value
  - In the case of continuous attributes it is likely that no specific numerical value will occur more than a small number of times. In this case the estimate used is generally the average value.

- **Reducing the Number of Attributes**
  - Feature reduction, dimension reduction

Repository of Datasets

- Most of the commercial datasets used by companies for data mining are unsurprisingly not available for others to use. However there are a number of ‘libraries’ of datasets that are readily available for downloading from the World Wide Web free of charge by anyone.
- The UCI Repository of Datasets
  - https://archive.ics.uci.edu/ml/index.php
- Kaggle
  - https://www.kaggle.com/

Classroom Exercises

1. What is the difference between labelled and unlabelled data? Explain with examples
2. The following information is held in an employee database. Name, Date of Birth, Sex, Weight, Height, Marital Status, Number of Children. What is the type of each variable?
3. Give two ways of dealing with missing data values.
4. Describe a scenario of your previous course project and explain each type of data that.

Assignment/Exercises

1. Propose an algorithm, in pseudocode or in your favorite programming language, for the following:
   - The automatic generation of a concept hierarchy for nominal data based on the number of distinct values of attributes in the given schema.
   - The automatic generation of a concept hierarchy for numeric data based on the *equal-width* partitioning rule.
   - The automatic generation of a concept hierarchy for numeric data based on the *equal-frequency* partitioning rule.

Lecture Reference

- Principles of Data Mining – Max Bramer (2nd Edition)
  - Chapter – 1 (*Data for Data Mining*)
- Data Mininig Concepts and Techniques – Jiawei Han, Michaline Kamber, Jian Pei, 3rd Ed.
- http://www.dictionary.com/browse/object
- https://docs.oracle.com/javase/tutorial/java/concepts/object.html

References

- Principles of Data Mining – Max Bramer (2nd Edition)
- Data Mininig Concepts and Techniques – Jiawei Han, Michaline Kamber, Jian Pei, 3rd Ed.
- Weka - open source machine learning software
  - https://www.cs.waikato.ac.nz/ml/weka/
- The UCI Repository of Datasets
  - https://archive.ics.uci.edu/ml/index.php
- Kaggle: Your Machine Learning and Data Science Community
  - https://www.kaggle.com/