Chapter 2:

Data Mining

What is Data Mining?

- **Data mining** is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data source.
- **Data mining** refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data.
- To make data mining more efficient, the data warehouse should have an aggregated or summarized collection of data.

Cont…

- data mining applications should be strongly considered early, during the design of a data warehouse.
- Also, data mining tools should be designed to facilitate their use in conjunction with data warehouses.
- In fact, for very large databases running into terabytes and even petabytes of data, successful use of data mining applications will depend first on the construction of a data warehouse.
- Focused on **hypothesis generation**, not on **hypothesis testing**

Alternative names of data mining

- knowledge extraction,
- Knowledge discovery(mining) from databases (KDD),
- data/pattern analysis,
- data archeology,
- information harvesting,
- business intelligence, etc

cont…

- Note that: query processing systems, Expert statistical data analysis or Information retrieval systems are not data mining tasks.
- The result of mining may be to discover the following type of new information:
- Association rules
- Sequential patterns
- Classification trees

Reporting the result of data mining

- The results of data mining may be reported in a variety of formats, such as
- listings,
- graphic outputs,
- summary tables,or
- visualizations.

Knowledge Discovery in Databases (KDD)

- typically encompasses more than data mining. The knowledge discovery process comprises six phases:
- data selection,
- data cleansing,
- enrichment,
- data transformation or encoding,
- data mining, and
- the reporting and
- display of the discovered information.

Statistics vs. Data Mining

Statistics	Data Mining
Confirmative	Explorative
Small data sets/File-based	Large data sets/Databases
Small number of variables	Large number of variables
Deductive	Inductive
Numeric data	Numeric and non-numeric (including txt, networks)
Clean data	Data cleaning

Goal of Data Mining

- Data mining is typically carried out with some end goals or applications
- these goals fall into the following classes:
- prediction
- identification
- classification, and
- optimization.

Application of Data Mining

- In particular, areas of significant payoffs are expected to include the following:
- Marketing.
- Finance.
- Manufacturing.
- Health Care.

Challenges in Data Mining

Some of the challenges with data mining are:

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge
- Pattern evaluation and knowledge integration
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Data source for DM applications

- Where are the data sources for analysis?
  - Credit card transactions,
  - loyalty cards,
  - discount coupons,
  - customer complaint calls,
  - Customer calls
  - Log files
  - Transaction files etc.
  - The best-known tool for data mining applications is **Weka**

Data Mining Functionalities

- Data mining task can be broadly classified into two as:
- Descriptive
- Predictive

different kinds of data mining functionalities

- Concept /class description: Characterization and discrimination
- Association Analysis
- Classification and prediction
- Clustering analysis
- Outlier analysis
- Evolution analysis

A Multi-Dimensional View of Data Mining Classification

- Different views, different classifications:
- Kinds of Databases to be mined
- Kinds of Knowledge to be mined
- Kinds of Techniques utilized
- Kinds of Applications adapted