Imp terms

Full forms

Module-1 : Intro to Data Mining

1.1. Data Mining Definations

Data mining(DM) also known as knowledge discovery from data (KDD). Here, in this course we will study data mining concepts & techniques for finding interesting patterns from data.

We live in a data age, where data is coming from everywhere and in gigantic volumes. Powerful & versatile tools are badly needed to automatically uncover valuable information from tremendous amounts of data to transform data into organized language.

Data Mining is the process of discovering interesting patterns knowledge from large amounts of data.

Data Warehouse is the repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. It includes data cleaning, data integration & OLAP (Online Analytical Processing)

is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. This process is discussed in Chapters 3 and 4. Figure 1.6 shows the typical framework for construction and use of a data warehouse for *AllElectronics*.

To facilitate decision making, the data in a data warehouse are organized around *major subjects* (e.g., customer, item, supplier, and activity). The data are stored to provide information from a *historical perspective*, such as in the past 6 to 12 months, and are typically *summarized*. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region.

A data warehouse is usually modeled by a multidimensional data structure, called a **data cube**, in which each **dimension** corresponds to an attribute or a set of attributes in the schema, and each **cell** stores the value of some aggregate measure such as *count* or *sum(sales_amount)*. A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

Note: A popular trend in IT industry is to perform data cleaning and integration as a preprocessing step, where the resulting data are stored in data warehouse.

What is (not) Data Mining?

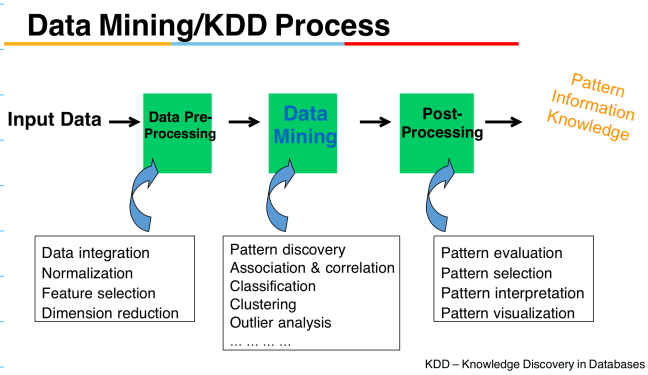
What is not Data Mining?

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

What is Data Mining?

- Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Some believe KDD & DM to be interchangable terms whereas others contest DM is merely a step in KDD Process.



KDD – Knowledge Discovery in Databases

KDD step is as follows: (CIS TM-PK)

1. Data Cleaning: to remove noise & inconsistent data
2. Data Integration: to combine multiple data sources
3. Data Selection: where relevant data in hand picked for analysis
4. Data Transformation: where data is transform using various techniques (summary or aggregation operations) to make it appropriate for mining
5. Data mining: a process where intelligent methods are applied to extract data patterns
6. Pattern Evaluation: identifying strictly increasing patterns representing knowledge based on different measures.
7. Knowledge Presentation: present mined knowledge using (visualization & knowledge representation techniques

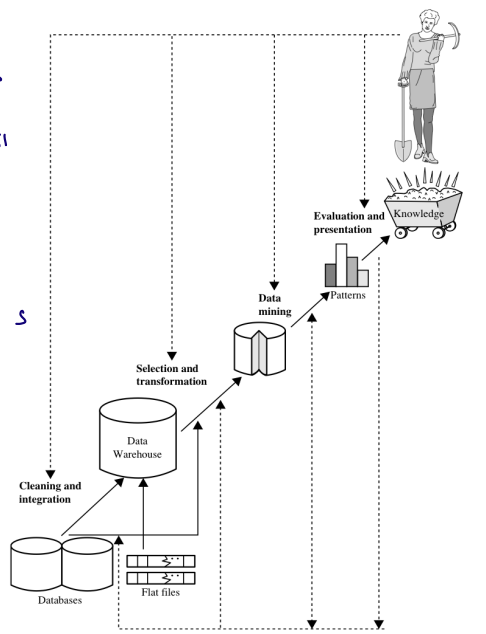


Figure 1.4 Data mining as a step in the process of knowledge discovery.

What kinds of data can be mined?

DM can be applied to any kind of data as long as the data are meaningful for the target applications.
Example : Database data, Data warehouse data, & transactional data.
Also, to data streams, ordered/sequenced data, graph and networked data, spatial data, text data, multimedia data & WWW.

Data Mining on Diverse kinds of Data

Besides relational database data (from operational or analytical systems) there are many other kinds of data that have diverse forms and structure and different semantic meanings.

Examples of data can be :

- time-related or sequence data (e.g., historical records, stock exchange data, and tir series and biological sequence data),
- data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
- spatial data (e.g., maps),
- engineering design data (e.g., the design of buildings, system components, or integrated circuits),
- hypertext and multimedia data (including text, image, video, and audio data),
- graph and networked data (e.g., social and information networks), and
- the Web (a widely distributed information repository).

Diversity of data brings in new challenges such as handling special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity)

By providing multidimensional data views and the precomputation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing operations make use of background knowledge regarding the domain of the data being studied to allow the presentation of data at *different levels of abstraction*. Such operations accommodate different user viewpoints. Examples of OLAP operations include **drill-down** and **roll-up**, which allow the user to view the data at differing degrees of summarization, as illustrated in Figure 1.7(b). For instance, we can drill down on sales data summarized by *quarter* to see data summarized by *month*. Similarly, we can roll up on sales data summarized by *city* to view data summarized by *country*.

Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis. **Multidimensional data mining** (also called **exploratory multidimensional data mining**) performs data mining in multidimensional space in an OLAP style. That is, it allows the exploration of multiple combinations of dimensions at varying levels of granularity in data mining, and thus has greater potential for discovering interesting patterns representing knowledge. An overview of data warehouse and OLAP technology is provided in Chapter 4.

Data Mining Functionalities

There are a number of ***data mining functionalities***. These include characterization and discrimination (Section 1.4.1); the mining of frequent patterns, associations, and correlations (Section 1.4.2); classification and regression (Section 1.4.3); clustering analysis (Section 1.4.4); and outlier analysis (Section 1.4.5). Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: **descriptive** and **predictive**. Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions.

Class/Concept description.

Let us assume, you have retail store data. Computer & printers are considered as classes of items. Now two concepts, bigspender, & budget spender can be considered as concepts of custom.
When you explain data with class/concepts, it comes as summarized, concise & precise. This way is called Class/concept description.

These descriptions can be derived as

① Characterization, by summarising the class data under study (target class)

or ② Discrimination, by comparing target class with one/set of comparative classes (contrasting class)

Example 1.5 **Data characterization.** A customer relationship manager at *AllElectronics* may order the following data mining task: *Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics.* The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings. The data mining system should allow the customer relationship manager to drill down on any dimension, such as on *occupation* to view these customers according to their type of employment. ■

Example 1.6 **Data discrimination.** A customer relationship manager at *AllElectronics* may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year). The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree. Drilling down on a dimension like *occupation*, or adding a new dimension like *income_level*, may help to find even more discriminative features between the two classes. ■

Frequent Patterns → patterns that occur frequenty

Kind

frequent itemset	frequence sequence/ sequential pattern	frequency substructures
refers to set of items that appear in transactional data together. Ex. Milk & Bread	eg. a person buy a laptop, followed by graphic card & then digital camera	

Example 1.7 **Association analysis.** Suppose that, as a marketing manager at *AllElectronics*, you want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the *AllElectronics* transactional database, is

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”}) \; [support = 1\%, confidence = 50\%],$$

where X is a variable representing a customer. A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as “*computer* $\Rightarrow$ *software* [1%, 50%].”

Suppose, instead, that we are given the *AllElectronics* relational database related to purchases. A data mining system may find association rules like

$$age(X, \text{“}20..29\text{”}) \wedge income(X, \text{“}40K..49K\text{”}) \Rightarrow buys(X, \text{“}laptop\text{”})$$
$$[support = 2\%, confidence = 60\%].$$

The rule indicates that of the *AllElectronics* customers under study, 2% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer) at *AllElectronics*. There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age*, *income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**. ■

Classification & Regression for Predictive Analysis

Classification: process of finding a model that describes & distinguishes data class / concept

→ classification predicts categorical (discrete & unordered) labels

Regression: predicts continuous valued functions

→ used to predict missing or unavailable numerical data values

→ Regression Analysis is a statistical methodology often used for numerical predictions

→ it also encompasses identification of distribution trends based on available data

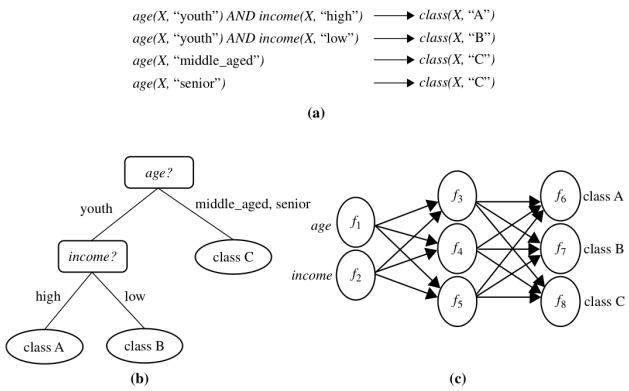


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Example 1.8 **Classification and regression.** Suppose as a sales manager of *AllElectronics* you want to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response* and *no response*. You want to derive a model for each of these three classes based on the descriptive features of the items, such as *price*, *brand*, *place_made*, *type*, and *category*. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

Suppose that the resulting classification is expressed as a decision tree. The decision tree, for instance, may identify *price* as being the single factor that best distinguishes the three classes. The tree may reveal that, in addition to *price*, other features that help to further distinguish objects of each class from one another include *brand* and *place_made*. Such a decision tree may help you understand the impact of the given sales campaign and design a more effective campaign in the future.

Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale at *AllElectronics*, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function (or ordered value.) ■

Classification: Application 1 (Use case)

Direct Marketing

- Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
- Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

Classification: Application 2

Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

Customer Attrition/Churn:

- Goal: To predict whether a customer is likely to be lost to a competitor.
- Approach:
  - Use detailed record of transactions with each of the past and present customers, to find attributes.
    - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal.
  - Find a model for loyalty.

Cluster Analysis

Clustering analyze data without using class label. In fact, clustering is used to generate class labels for group of data.
The objects are based on the principles of

① maximizing the intraclass similarity
shows the distance b/w data point of one cluster with the other data point in other cluster

& ② minimising the interclass similarity
shows the distance b/w data points with cluster centre

Clusters are formed in such way that cluster have high similarity in comparison to one another & quite dissimilar to objects in another cluster.
Each cluster formed can be viewed as a class of objects from which rules can be defined.

Clustering: Application 1

Market Segmentation:

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

Document Clustering:

- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Outlier Analysis / Anomaly detection

Objects that do not comply with general behaviour or model of the data are called outliers.
Most of the time outliers are disregared considering them as noise or exception.
In some cases, outliers are the heroes
eg. Fraudlent usage of credit cards.

Data Preparation

Data needs to be understood. It requires descriptive statistics such as mean, median, mode, standard deviation, and range for each attribute

Data quality is an ongoing concern wherever data is collected, processed, and stored.

- The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.
- it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models to ensure a certain degree of data quality

Missing Values

- Need to track the data lineage of the data source to find right solution

Data Types and Conversion

- The attributes in a data set can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical
- data mining algorithms impose different restrictions on what data types they accept as inputs

Transformation

- Can go beyond type conversion, may include dimensionality reduction or numerosity reduction

Outliers are anomalies in the data set

- May occur legitimately or erroneously.

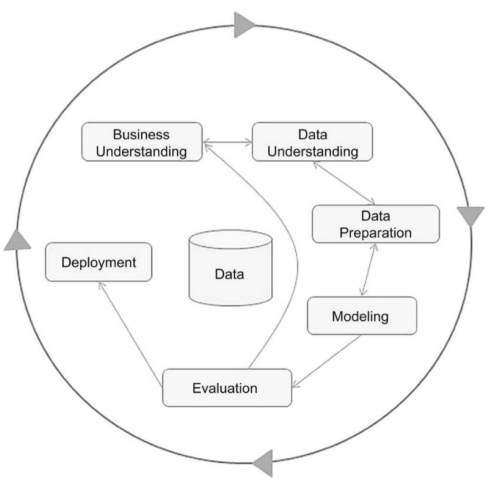
Feature Selection

- Many data mining problems involve a data set with hundreds to thousands of attributes, most of which may not be helpful. Some attributes may be correlated, e.g. sales amount and tax .

Data Sampling may be adequate in many cases

innovate achieve

CRISP data mining framework



CRISP is the most popular methodology for analytics, data mining, and data science projects, with 43% share as per 2014 KDnuggets Poll.

CRISP-DM was conceived in 1996. In 1997 it got underway as a European Union project, led by SPSS, Teradata, Daimler AG, NCR Corporation and OHRA.

Data Science Process Alliance

What is CRISP DM?

CRISP-DM Diagram. Inspired by Wikipedia

The CRoss Industry Standard Process for Data Mining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

Published in 1999 to standardize

Challenges of Data Mining

Scalability Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive data sets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

High Dimensionality It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases.

Heterogeneous and Complex Data Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyperlinks; DNA data with sequential and three-dimensional structure; and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on the Earth's surface. Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structured text and XML documents.

Data Ownership and Distribution Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. Among the key challenges faced by distributed data mining algorithms include (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security issues.

Non-traditional Analysis The traditional statistical approach is based on a hypothesize-and-test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically not the result of a carefully designed experiment and often represent opportunistic samples of the data, rather than random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.

Gartner Magic Quadrant

The Magic Quadrant: Your view into smarter technology purchases

How does a Gartner Magic Quadrant work?

A Magic Quadrant is a tool that provides a graphical competitive positioning of technology providers to help you make smart investment decisions. Thanks to a uniform set of evaluation criteria, a Magic Quadrant provides a view of the four types of technology providers in any given field:

Leaders execute well against their current vision for changing market rules, but do not yet execute well.

Visionaries understand where the market is going or have a vision for changing market rules, but do not yet execute well.

Niche Players focus successfully on a small segment, or are unfocused and do not out-innovate or outperform others.

Challengers execute well today or may dominate a large segment, but do not demonstrate an understanding of market direction.

The Magic Quadrant is more than just a diagram. You also get:

Custom category weighting

Weight critical vendor categories to get insights tailored to your business priorities.

Historical perspective

See how the vendor space has evolved through the years to better understand overall trends and upstarts.

User reviews

Read what your peers have to say about the solutions they've implemented.