
Data Mining

Data Mining

- Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data.
- Data mining is also called ***Knowledge Discovery in Database (KDD)***.

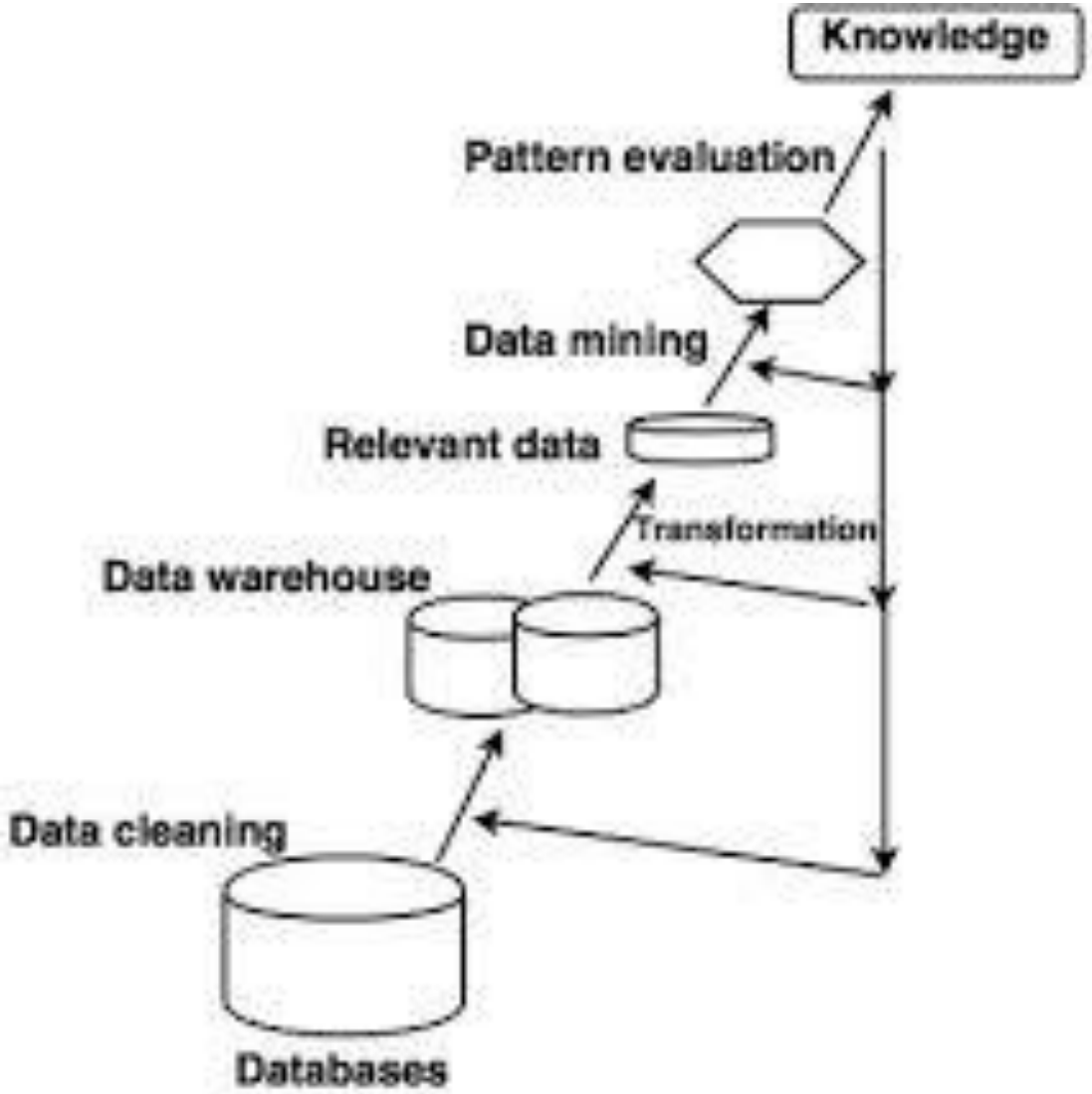
Definition

Data mining is the process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data Mining

The knowledge discovery process includes:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation.



Knowledge Discovery in Databases (KDD)

Data Mining – Other Definitions

1. Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.
2. Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures.
3. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.
4. Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Mining vs Data Science

- Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective.
- This process includes various types of services such as:
  - text mining,
  - web mining,
  - audio and video mining,
  - pictorial data mining, and
  - social media mining
- This is usually done through software that is simple or highly specific.

Advantages of Data Mining

1. The Data Mining technique enables organizations to obtain knowledge-based data.
2. Data mining enables organizations to make lucrative modifications in operation and production.
3. Compared with other statistical data applications, data mining is a cost-efficient.
4. Data Mining helps the decision-making process of an organization.
5. It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
6. It can be induced in the new system as well as the existing platforms.
7. It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

1. There is a probability that the organizations may sell useful data of customers to other organizations for money.

   - As per the report, American Express has sold credit card purchases of their customers to other organizations.

2. Many data mining analytics software is difficult to operate and needs advance training to work on.
3. Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
4. The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications

Data Mining Applications

- Data Mining is primarily used by organizations with intense consumer demands-
  - Retail,
  - Communication,
  - Financial,
  - marketing company,
  - determine price,
  - consumer preferences,
  - product positioning, and
  - impact on sales, customer satisfaction, and corporate profits.
- Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.

Data Mining Techniques

- There are four main techniques
  - Predictive Modelling
  - Database Segmentation
  - Link Analysis
  - Deviation Direction
- Many applications may work well when several or a combination of operations are used

Data Mining Techniques

1. Predictive Modelling

- This technique uses observations to form a model of the important characteristics of some phenomenon
- This technique can be used to analyse an existing database to determine some essential characteristics about the data set.
- Uses supervised learning.

- There are two main Techniques used in predictive modelling:
  - Classification
  - Regression

Data Mining Techniques

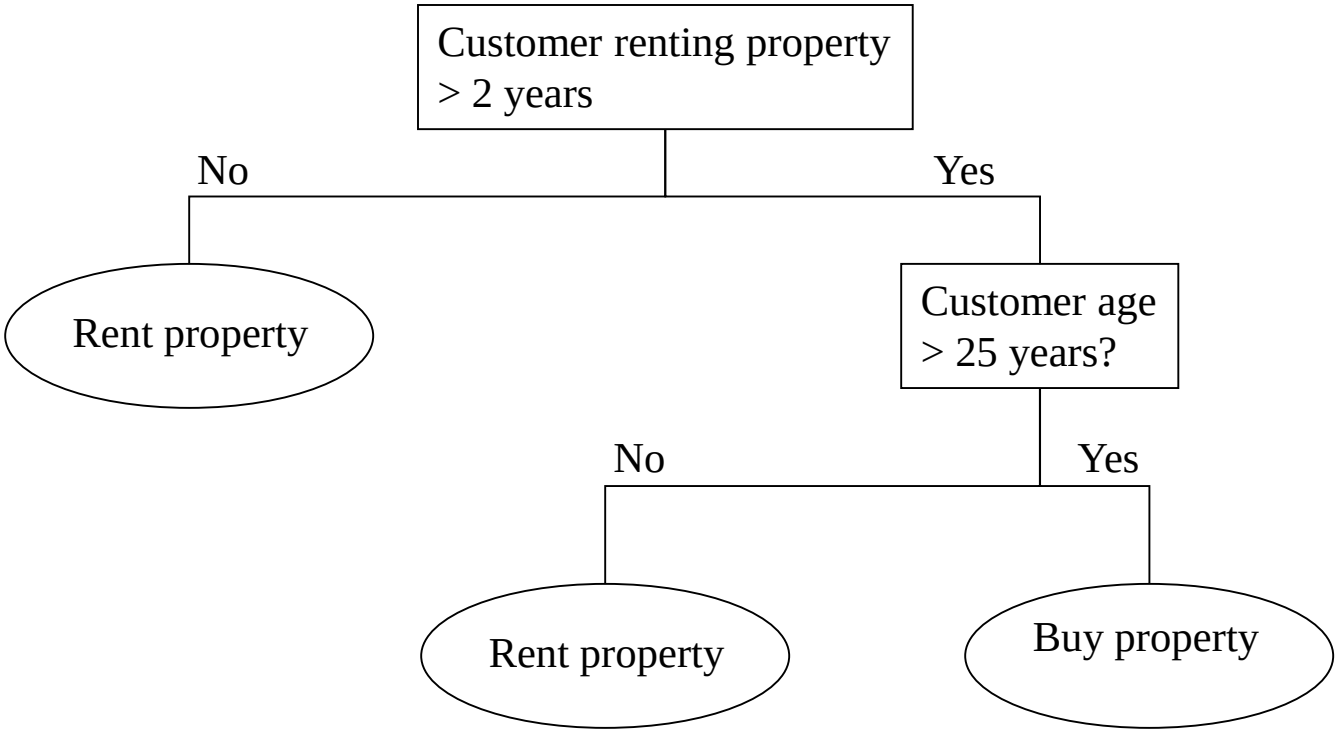
a. Classification

- It is used to establish a specific predetermined class for each record in a database from a finite set of possible class values, e.g. if a customer has rented for > 2 years and > 25 years old then they are most likely to buy property.
- Can use the following classifiers: neural network, decision tree, Bayes Naïve etc

b. Value prediction (Regression)

- It is used to estimate a continuous numeric value that is associated with a database record,
- It uses statistical techniques e.g. linear/non-linear regression.

Classification Example- Tree Induction



Source: Connolly and Begg

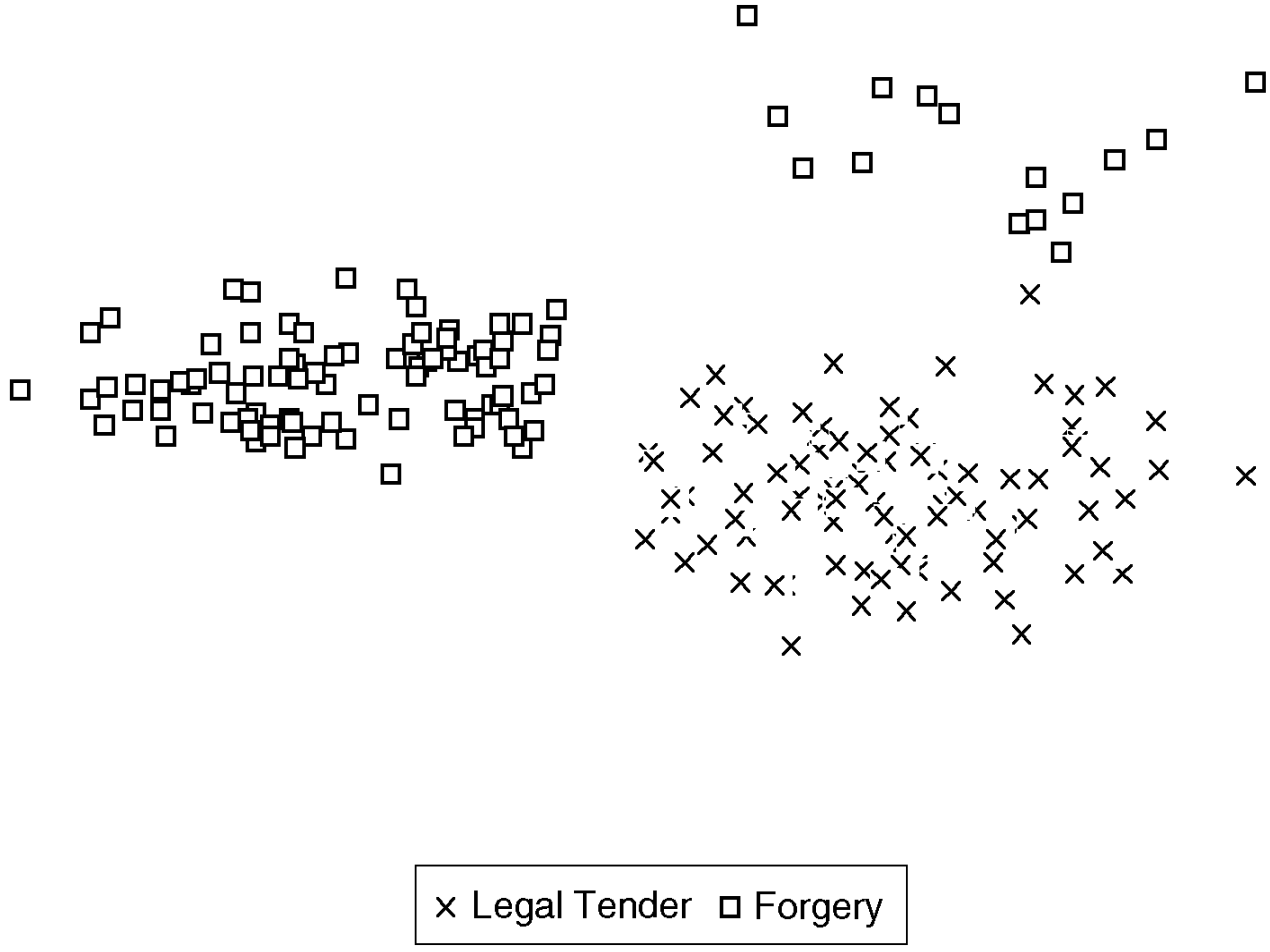
Data Mining Techniques

2. Database Segmentation (Cluster Analysis)

- This techniques creates clusters by partitioning a database into an unknown number of segments (or clusters) of records which share a number of propertiesi.e. homogenous
- Uses unsupervised learning to discover sub-populations in the database.

- The two main Techniques in database segmentation are:
  - Demographic clustering
  - Neural clustering

Segmentation: Scatterplot Example



Source: Connolly and Begg

Data Mining Techniques

3. Link Analysis (Association Rule Analysis)

- This technique is used to establish associations between individual records (or sets of records) in a database
  - e.g. 'when a customer rents property for more than two years and is more than 25 years old, then in 40% of cases, the customer will buy the property'

- The main Techniques used in link analysis are:
  - Association discovery
  - Sequential pattern discovery
  - Similar time sequence discovery

Data Mining Techniques

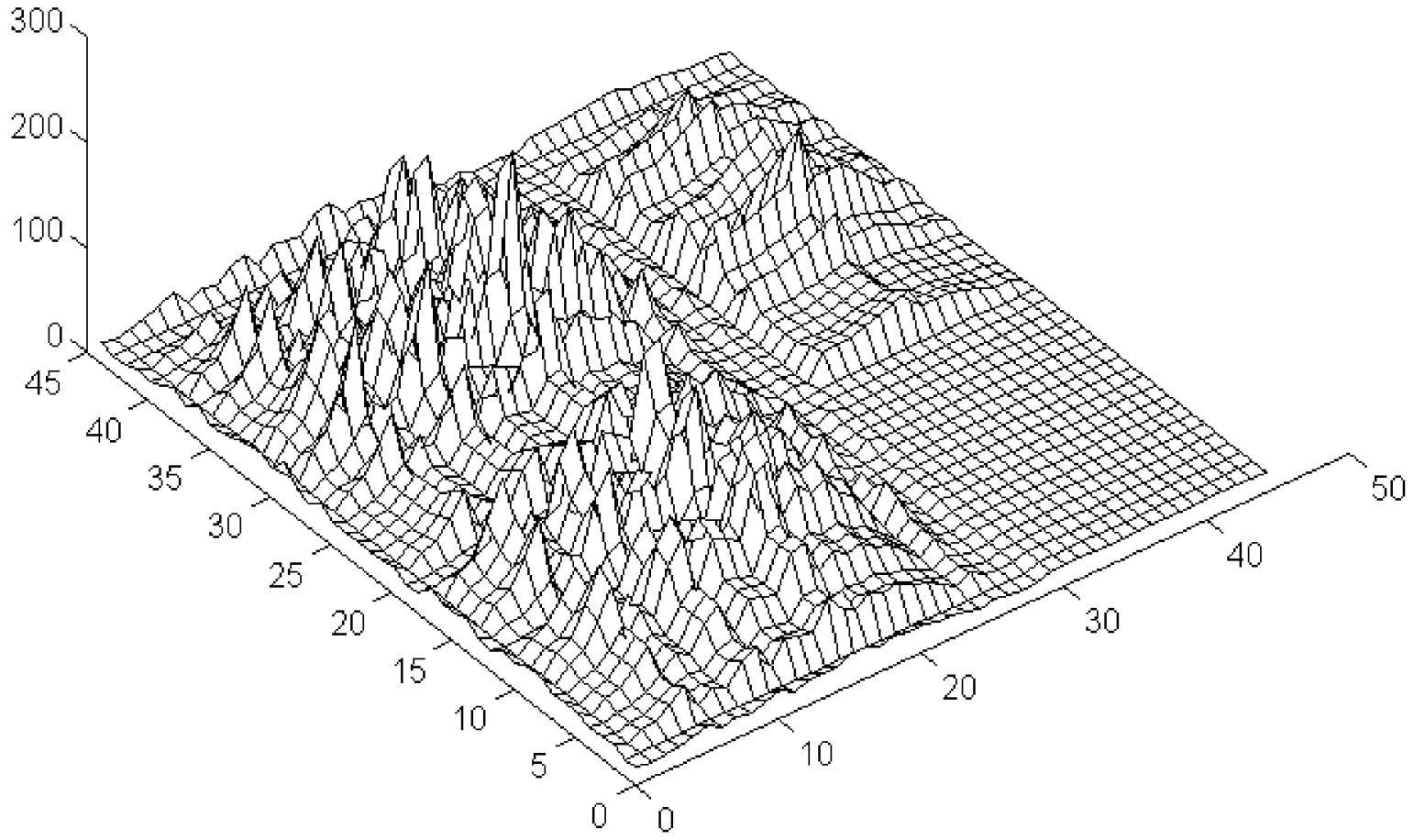
a. Association discovery – items which imply the presence of other item in same event
a. Sequential discovery – presence of 1 set of item implies presence of another in a period of time (e.g. long term customer buying behaviour)
b. Similar time sequence discovery – discovery of link between 2 sets of data that are time dependent, e.g. buying property -> buy household goods within 2 months.

Data Mining Techniques

4. Deviation Detection

- This technique is used to identifies records or 'outliers', which deviates from some known expectation or norm (value that are out of the ordinary)
- Can be done either statistically (e.g. linear regression) or by visualisation (e.g. graphically).

Deviation Detection: Visualisation Example



Source: Connolly and Begg

Mining and Warehousing

- Data warehouse is the ideal data source for data mining.
- Data mining needs single, separate, clean, integrated, self-consistent data source:
  - It is populated with clean, consistent data
  - Contains multiple sources that allow to discover as many inter-relationships as possible
  - Utilises Query capabilities that allow for selection of relevant subsets of records and fields
  - Has capability to go back to data source i.e. provides a way for data mining results to allow further investigation of uncovered patterns.

Further Reading

- Connolly and Begg, chapters 31 to 34.
- W H Inmon, *Building the Data Warehouse*, New York, Wiley and Sons, 1993.
- Benyon-Davies P, Database Systems (2nd ed), Macmillan Press, 2000, ch 34, 35 & 36.