Data Mining

UNIT - I: Data Mining Basics

Introduction: Definition of data mining - data mining vs. query tools - machine learning - steps in data mining process - overview of data mining techniques.

UNIT - II: Data Models

Multidimensional Data Model - Data Cube - Dimension Modeling - OLAP Operations - Meta Data - Types of Meta Data.

UNIT - III: Data Editing

Data Pre-Processing and Characterization: Data Cleaning - Data Integration and Transformation - Data Reduction - Data Mining Query Language - Generalization - Summarization - Association Rule Mining

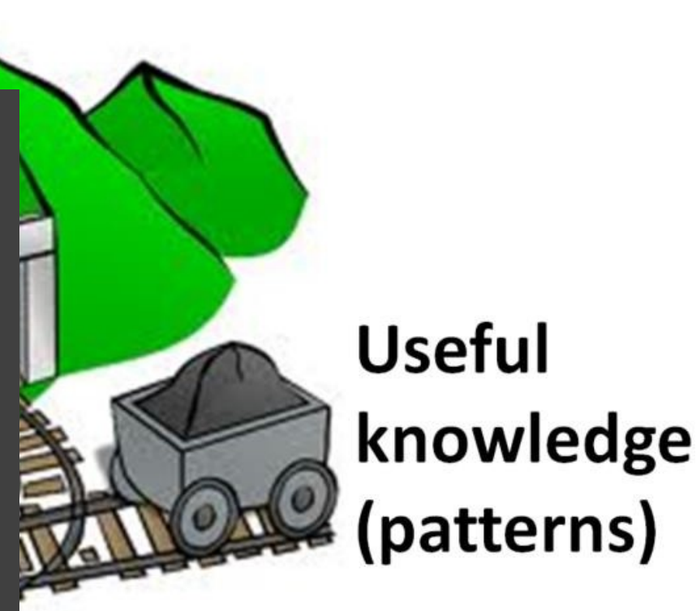
UNIT - IV: Classification

Classification: Classification - Decision Tree Induction - Bayesian Classification - Prediction - Back Propagation - Cluster Analysis - Hierarchical Method - Density Based Method - Grid Based Method - Outlier Analysis.

UNIT - V: Analysis

Cluster analysis: Types of data - Clustering Methods - Partitioning methods - Model based clustering methods - outlier analysis. Advanced topics: Web Mining - Web Content Mining - Structure and Usage Mining - Spatial Mining - Time Series and Sequence Mining.

n of
ing'



Definition of Data Mining:

- In simple words, data mining is defined as a process used to extract valuable information from a larger set of any raw data.
- Data mining has applications in multiple fields, like science, medical, banking, online shopping and research.
- As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions effectively.

- Data Mining helps businesses be closer to their objective and make better decisions.
- **Data mining** is a field of research that has emerged in the 1990s.
- It is very popular today, sometimes under different names such as “**big data**” and “**data science**“, which have a similar meaning.
- To give a short definition of **data mining,** it can be defined as a set of techniques for automatically analyzing data to discover interesting knowledge or pasterns in the data.

The reasons why data mining has become popular

- The reasons why data mining has become popular is that storing data electronically has become very cheap and that transferring data can now be done very quickly and easily.
- Thus, many organizations now have **huge amounts of data** stored in databases, that needs to be analyzed.
- To address this problem, automatic techniques have been designed to **analyze data and extract interesting patterns, trends or other useful information**. This is the purpose of **data mining**.

- In general, data mining techniques are designed either to **explain or understand the past.**
  - **predict the future.**
  - **Example:** why a plane has crashed.
    - Predict if there will be an earthquake tomorrow at a given location.
- Data mining techniques are used to take decisions based on facts rather than intuition.

What are the applications of data mining?

There is a wide range of data mining techniques (algorithms), which can be applied in all kinds of domains where data has to be analyzed. Some example of data mining applications are:

- **fraud detection,**
- **stock market price prediction,**
- **Analyzing the behavior of customers** in terms of **what they buy**

In general data mining techniques are chosen based on:

- the **type of data to be analyzed,**
- the type of **knowledge or patterns to be extracted from the data**,
- how the **knowledge will be used**.

Data mining Vs Query Tools

1. Users who are inclined toward statistics use Data Mining.
2. They utilize statistical models to look for hidden patterns in data.
3. Data miners are interested in finding useful relationships between different data elements, which is ultimately profitable for businesses.
4. Data mining is also known as Knowledge Discovery in Data (KDD).
5. As mentioned above, it is a field of computer science, which deals with the extraction of previously unknown and interesting information from raw data.
6. Due to the exponential growth of data, especially in areas such as business, data mining has become very important tool to convert this large wealth of data into business intelligence.
7. as manual extraction of patterns has become seemingly impossible in the past few decades.

- For example, it is currently been used for various applications such as social network analysis, fraud detection and marketing.
- Data mining usually deals with following four tasks: clustering, classification, regression, and association.
- Clustering is identifying similar groups from unstructured data.
- Classification is learning rules that can be applied to new data and will typically include following steps: preprocessing of data, designing modeling, learning/feature selection and Evaluation/validation.
- Regression is finding functions with minimal error to model data.
- And association is looking for relationships between variables.
- Data mining is usually used to answer questions like what are the main products that might help to obtain high profit next year in Wal-Mart?

Query Tools

1. Query Tools are tools that help to analyze the data in a database.
2. Usually these query tools have a GUI front end with convenient ways to input queries as a set of attributes.
3. Once these inputs are provided the tool generates actual queries made up of the underlying query language used by the database.
4. SQL, T-SQL and PL/SQL are examples of query languages used in many popular databases today.
5. Then, these generated queries are executed against the databases and the results of the queries are presented or reported to the user in an organized and clear manner.
6. Typically, the user does not need to know a database-specific query language to use a Query tool.
7. Key features of Query tools are integrated query builder and editor, summery reports and figures, import and export features and advanced find/search capabilities.

What is the difference between Data mining and Query Tools?

1. Query tools can be used to easily build and input queries to databases.
2. Query tools make it very easy to build queries without even having to learn a database-specific query language.
3. Data Mining is a technique or a concept in computer science, which deals with extracting useful and previously unknown information from raw data.
4. Most of the times, these raw data are stored in very large databases.
5. Therefore Data miners can use the existing functionalities of Query Tools to preprocess raw data before the Data mining process.
6. **However, the main difference between Data mining techniques and using Query tools is that, in order to use Query tools the users need to know exactly what they are looking for, while data mining is used mostly when the user has a idea about what they are looking for.**

Machine Learning

- Machine learning is an application of artificial intelligence (AI)
- It provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.
- The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.

- **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.
- The intelligent systems built on machine learning algorithms have the capability to learn from past experience or historical data.
- Machine learning applications provide results on the basis of past experience.
- 10 real-life examples of how machine learning is helping in creating better technology to power today's ideas.

Machine Learning Applications

Image Recognition

- Image recognition is one of the most common uses of machine learning.

Speech Recognition

- Speech recognition is the translation of spoken words into the text. It is also known as computer speech recognition or automatic speech recognition.

Medical diagnosis

- Machine learning can be used in the techniques and tools that can help in the diagnosis of diseases.

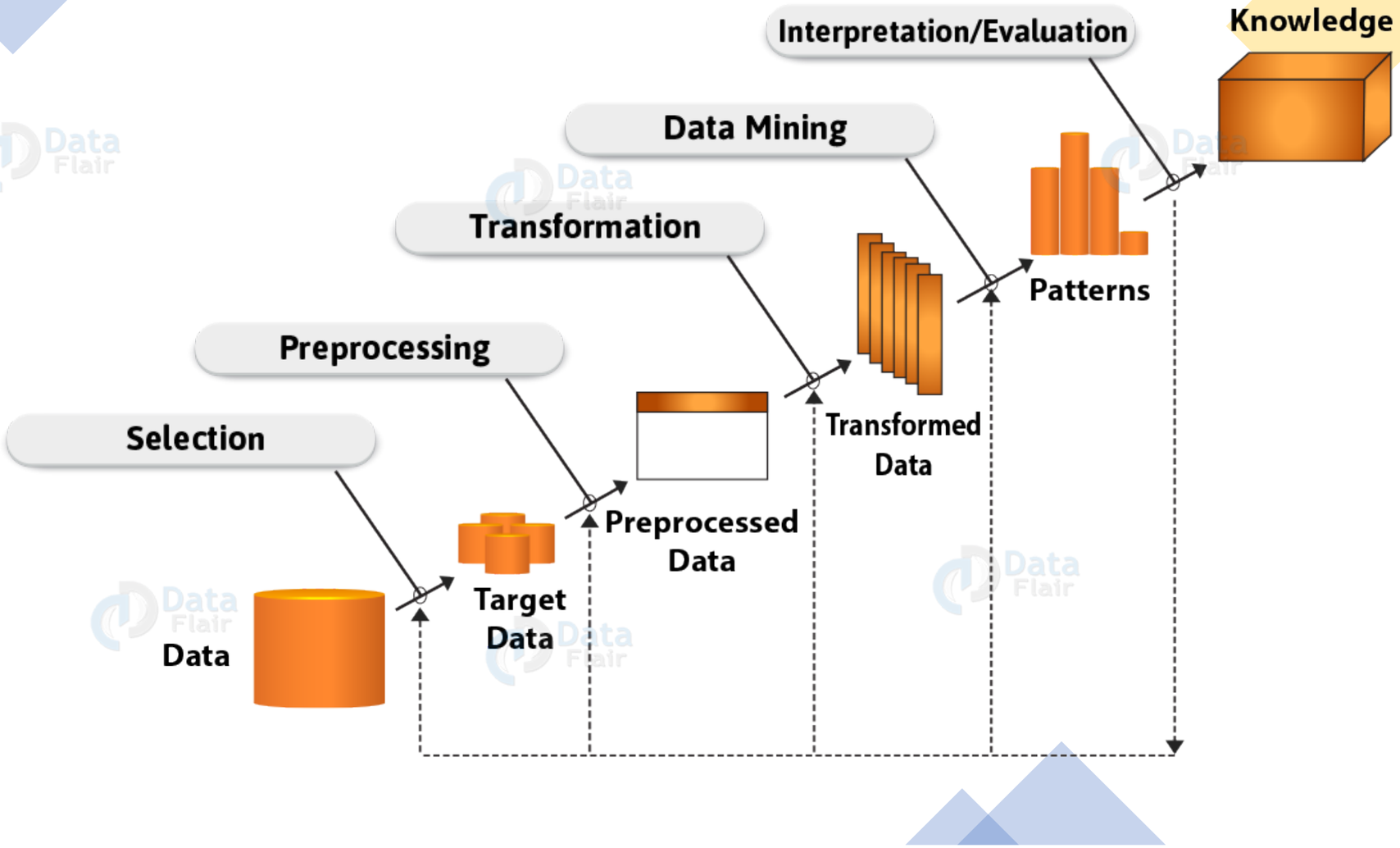
Classification

- A classification is a process of placing each individual under study in many classes.

Prediction

- Machine learning can also be used in the prediction systems.

steps in data mining process



Various Stages of Data Mining

To perform data mining, a process consisting of seven steps is usually followed. This process is often called the "**Knowledge Discovery in Database**" (**KDD**) process or **Various stages of Data Mining.**

1. **Data cleaning**: This step consists of cleaning the data by **removing noise or other inconsistencies** that could be a problem for analyzing the data.

2. **Data integration**: This step consists of integrating data **from various sources to prepare the data that needs to be analyzed**. For example, if the data is stored in multiple databases or file, it may be necessary to integrate the data into a single file or database to analyze it.

3. **Data selection**: This step consists of **selecting the relevant data** for the analysis to be performed.

4. **Data transformation**: This step consists of **transforming the data to a proper format** that can be analyzed using data mining techniques. For example, some data **mining techniques require that all numerical values are normalized**.

5. **Data mining**: This step consists of **applying some data mining techniques (algorithms) to analyze the data and discover interesting patterns** or extract interesting knowledge from this data.

6. **Evaluating the knowledge that has been discovered**: This step consists of evaluating the **knowledge that has been extracted from the data.** This can be done in terms of objective and/or subjective measures.

7. **Visualization**: Finally, the last step is to visualize the knowledge that has been extracted from the data.

Data Mining Techniques

1. Classification
2. Association
3. Clustering
4. Prediction
5. Decision trees

Classification

- Classification data mining techniques involve analyzing the various attributes associated with different types of data.
- Once organizations identify the main characteristics of these data types, organizations can categorize or classify related data.
- Doing so is critical for identifying, for example, personally identifiable information organizations may want to protect or redact from documents.
- Classification is a data mining function that assigns items in a collection to target categories or classes.
- The goal of classification is to accurately predict the target class for each case in the data.
- For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Association

- Association is a data mining technique related to statistics.
- It indicates that certain data (or events found in data) are linked to other data or data-driven events.
- It is similar to the notion of co-occurrence in machine learning, in which the likelihood of one data-driven event is indicated by the presence of another.
- The statistical concept of correlation is also similar to the notion of association.
- This means that the analysis of data shows that there is a relationship between two data events: such as the fact that the purchase of hamburgers is frequently accompanied by that of French fries.

Clustering

- ❖ Clustering is an analytics technique that relies on visual approaches to understanding data.
- ❖ Clustering mechanisms use graphics to show where the distribution of data is in relation to different types of metrics.
- ❖ Clustering techniques also use different colors to show the distribution of data.
- ❖ Graph approaches are ideal for using cluster analytics.
- ❖ With graphs and clustering in particular, users can visually see how data is distributed to identify trends that are relevant to their business objectives.

Prediction

- Prediction is a very powerful aspect of data mining that represents one of four branches of analytics.
- Predictive analytics use patterns found in current or historical data to extend them into the future.
- It gives organizations insight into what trends will happen next in their data.
- There are several different approaches to using predictive analytics.
- Some of the more advanced involve aspects of machine learning and artificial intelligence.

Decision trees

- Decision trees are a specific type of predictive model that lets organizations effectively mine data.
- Technically, a decision tree is part of machine learning, but it is more popularly known as a white box machine learning technique because of its extremely straightforward nature.
- A decision tree enables users to clearly understand how the data inputs affect the outputs. When various decision tree models are combined they create predictive analytics models known as a random forest.
- Complicated random forest models are considered black box machine learning techniques, because it's not always easy to understand their outputs based on their inputs.
- In most cases, however, this basic form of ensemble modeling is more accurate than using decision trees on their own.

Dimensional Modeling:

- **DIMENSIONAL MODELING (DM)** is a data structure technique optimized for data storage in a Data warehouse.
- The purpose of dimensional model is to optimize the database for fast retrieval of data.
- The concept of Dimensional Modelling was developed by Ralph Kimball and consists of "fact" and "dimension" tables.
- A Dimensional model is designed to read, summarize, analyze numeric information like values, balances, counts, weights, etc. in a data warehouse.
- In contrast, relation models are optimized for addition, updating and deletion of data in a real-time Online Transaction System.
- These dimensional and relational models have their unique way of data storage that has specific advantages.

- For instance, in the relational mode, normalization and ER models reduce redundancy in data. On the contrary, dimensional model arranges data in such a way that it is easier to retrieve information and generate reports.
- Hence, Dimensional models are used in data warehouse systems and not a good fit for relational systems.

- Elements of Dimensional Data Model
- Fact
- Dimension
- Attributes
- Fact Table
- Dimension table

Elements of Dimensional Data Model

Fact

Facts are the measurements/metrics or facts from your business process. For a Sales business process, a measurement would be quarterly sales number

Dimension

Dimension provides the context surrounding a business process event. In simple terms, they give who, what, where of a fact. In the Sales business process, for the fact quarterly sales number, dimensions would be

•Who – Customer Names Where – Location What – Product Name

In other words, a dimension is a window to view information in the facts.

Attributes

The Attributes are the various characteristics of the dimension.

In the Location dimension, the attributes can be

- State
- Country
- Zipcode etc.

Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes

Fact Table

A fact table is a primary table in a dimensional model.

A Fact Table contains

1. Measurements/facts
2. Foreign key to dimension table

Dimension table

- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- Dimension tables are de-normalized tables.
- The Dimension Attributes are the various columns in a dimension table
- Dimensions offers descriptive characteristics of the facts with the help of their attributes
- No set limit set for given for number of dimensions
- The dimension can also contain one or more hierarchical relationships

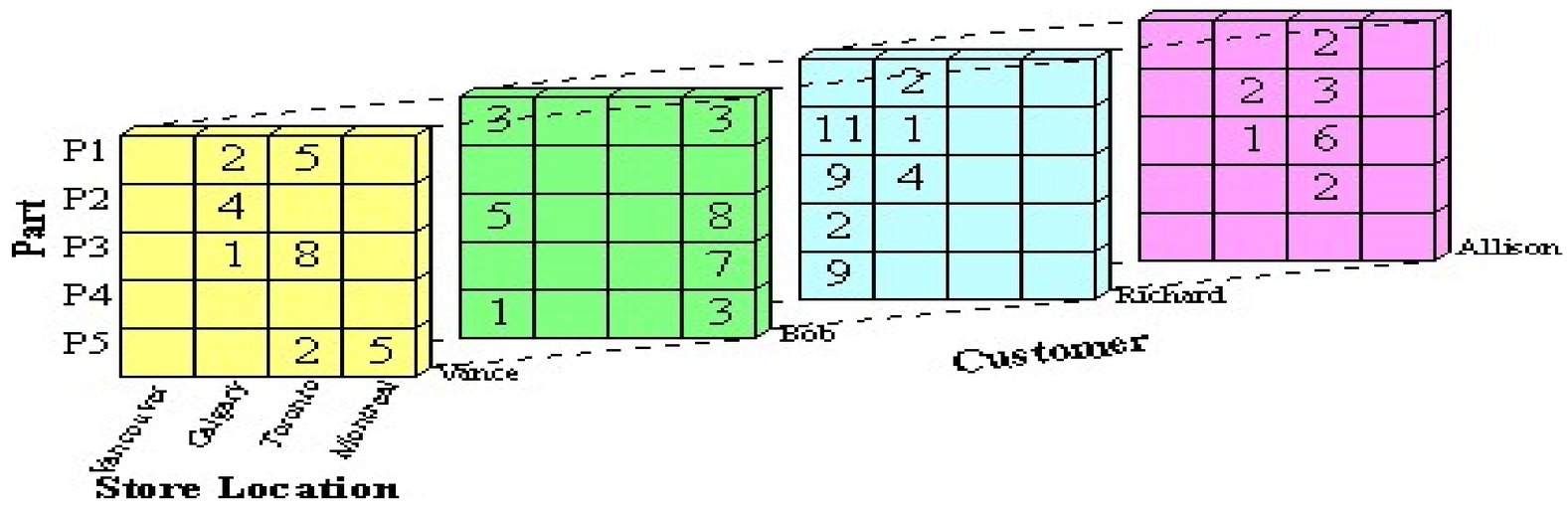
Data Cube

- A data cube refers is a three-dimensional (3D) (or higher) range of values that are generally used to explain the time sequence of an image's data.
- It is a data abstraction to evaluate aggregated data from a variety of viewpoints.
- A data cube can also be described as the multidimensional extensions of two-dimensional tables.
- It can be viewed as a collection of identical 3-D tables stacked upon one another.
- Data cubes are used to represent data that is too complex to be described by a table of columns and rows.
- As such, data cubes can go far beyond 3-D to include many more dimensions.

- A data cube is generally used to easily interpret data.
- It is especially useful when representing data together with dimensions as certain measures of business requirements.
- A cube's every dimension represents certain characteristic of the database.
- For example, daily, monthly or yearly sales.
- The data included inside a data cube makes it possible analyze almost all the figures for virtually any or all customers, sales agents, products, and much more.
- Thus, a data cube can help to establish trends and analyze performance.

- Data cubes are mainly categorized into two categories:
  - Multidimensional Data Cube:
    - Most OLAP products are developed based on a structure where the cube is patterned as a multidimensional array.
    - These multidimensional OLAP (MOLAP) products usually offers improved performance when compared to other approaches mainly because they can be indexed directly into the structure of the data cube to gather subsets of data.
    - When the number of dimensions is greater.
    - Compression techniques might help; however, their use can damage the natural indexing of MOLAP.

- Relational OLAP:
  - Relational OLAP make use of the relational database model.
  - The ROLAP data cube is employed as a bunch of relational tables compared to a multidimensional array.
  - Each one of these tables, known as a cuboid, signifies a specific view.



OLAP Operations

Meta data

Meta data

- Metadata is data that describes other data.
- Meta is a prefix that in most information technology usages means "an underlying definition or description."
- Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier.
- For example, *author*, *date created*, *date modified* and *file size* are examples of very basic document metadata.
- Having the ability to filter through that metadata makes it much easier for someone to locate a specific document.
- In addition to document files, metadata is used for:

Images Videos Spreadsheets Web pages

Types of Metadata

Metadata comes in several types and is used for a variety of broad purposes that can be roughly categorized as a business, technical, or operational.

•**Descriptive** metadata properties include title, subject, genre, author, and creation date, for example.

•**Rights** metadata might include copyright status, rights holder, or license terms.

•**Technical** metadata properties include file types, size, creation date and time, and type of compression. Technical metadata is often used for digital object management and interoperability.

- **Preservation** metadata is used in navigation. Example preservation metadata properties include an item's place in a hierarchy or sequence.
- **Markup languages** include metadata used for navigation and interoperability. Properties might include heading, name, date, list, and paragraph.

Descriptive Metadata

- Descriptive metadata is essential for discovering and identifying assets.
- Why? It's information that describes the asset, such as the asset's title, author, and relevant keywords.
- Descriptive metadata is what allows you to locate a book in a particular genre published after 2016, for instance, as a book's metadata would include both genre and publication date.
- In fact, the ISBN system is a good example of an early effort to use metadata to centralize information and make it easier to locate resources (in this case, books in a traditional library).

UNIT - III: Data Editing

Data Pre-Processing and Characterization:

Data Cleaning in Data Mining

- Quality of your data is critical in getting to final analysis.
- Any data which tend to be incomplete, noisy and inconsistent can effect your result.
- Data cleaning in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database.
- Data cleaning is one of the important parts of machine learning.
- However, proper data cleaning can make or break your project. Professional data scientists usually spend a very large portion of their time on this step.
- If we have a well-cleaned dataset, we can get desired results even with a very simple algorithm, which can prove very beneficial at times.

Some data cleaning methods :-

1. You can ignore the tuple. This is done when class label is missing. This method is not very effective, unless the tuple contains several attributes with missing values.

2. You can fill in the missing value manually. This approach is effective on small data set with some missing values.

3. You can replace all missing attribute values with global constant, such as a label like "Unknown" or minus infinity.

4. You can use the attribute mean to fill in the missing value.

5. For example customer average income is 25000 then you can use this value to replace missing value for income.

5 Use the most probable value to fill in the missing value.

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values

Invalid values

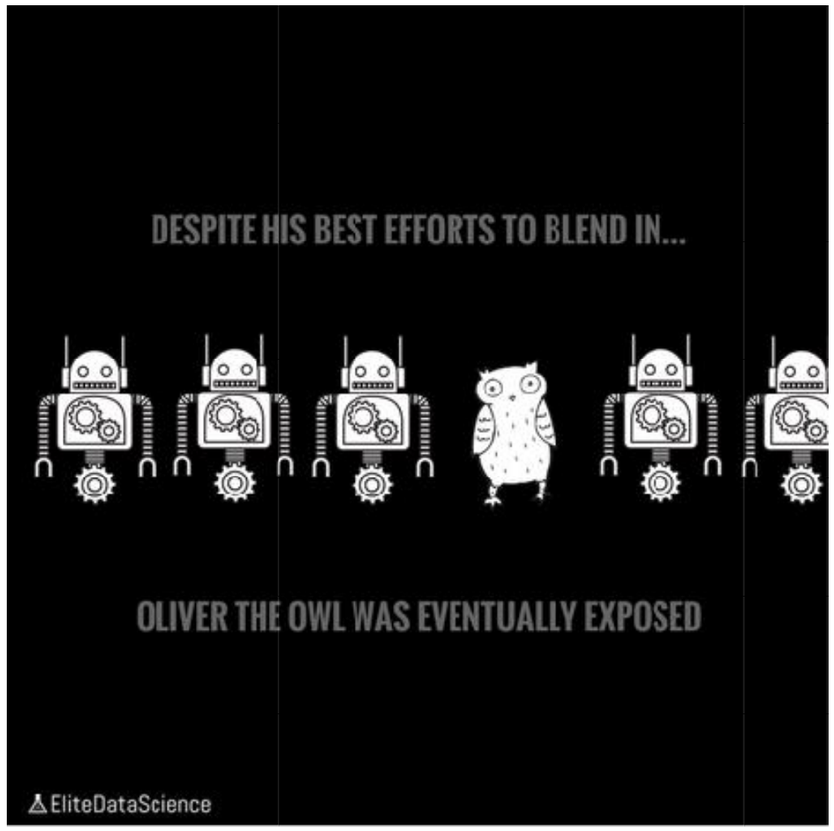
Misfielded values

Misspellings

Uniqueness

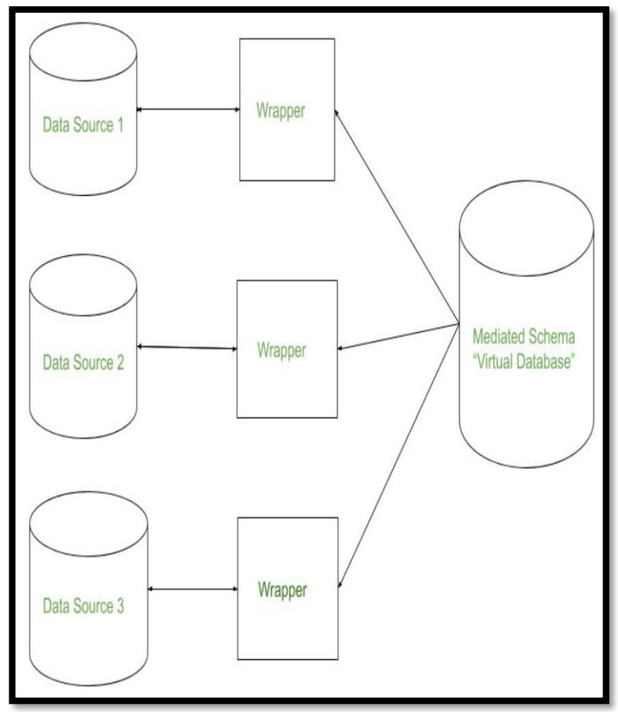
Formats

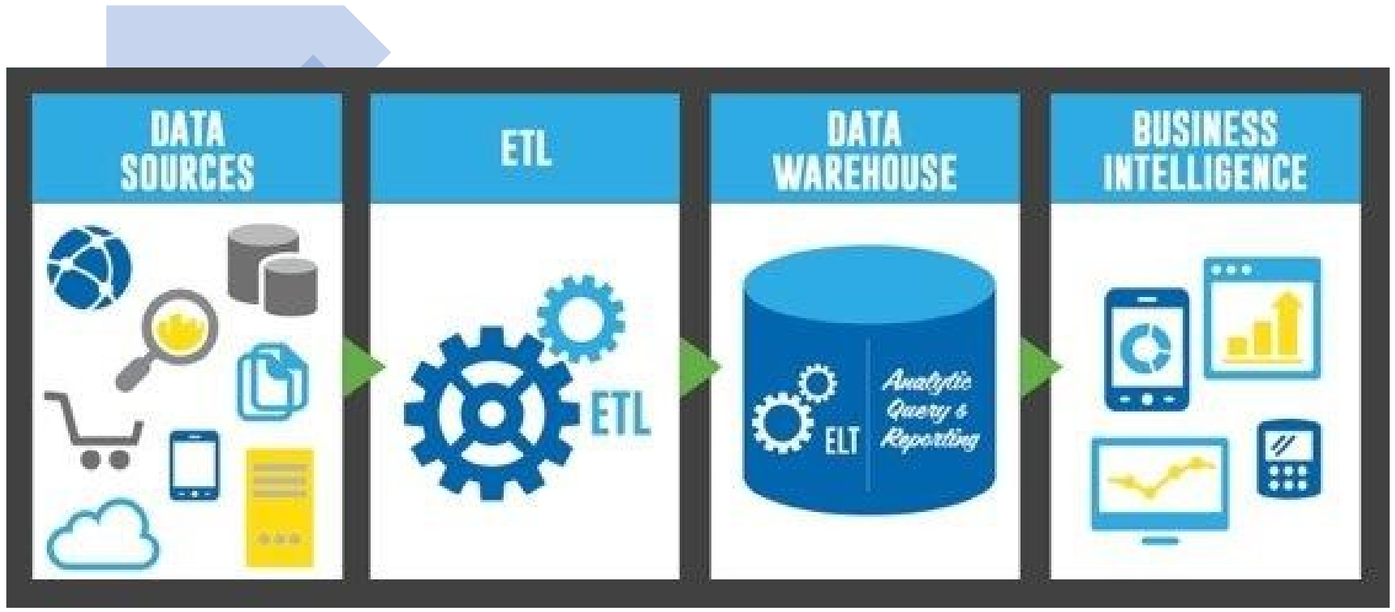
Attribute dependencies


DESPITE HIS BEST EFFORTS TO BLEND IN...
OLIVER THE OWL WAS EVENTUALLY EXPOSED
EliteDataScience

Data Integration

- Data Integration is a data preprocessing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data.
- These sources may include multiple data cubes, databases or flat files.
- The data integration approach are formally defined as triple <G, S, M> where,
    - G stand for the global schema,
    - S stand for heterogenous source of schema,
    - M stand for mapping between the queries of source and global schema.




DATA SOURCES
ETL
DATA WAREHOUSE
BUSINESS INTELLIGENCE
ETL
ELT
Analytic Query & Reporting

There are mainly 2 major approaches for data integration – one is “tight coupling approach” and another is “loose coupling approach”.

Tight Coupling:

- Here, a data warehouse is treated as an information retrieval component.
- In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation and Loading.

Loose Coupling:

•Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.

•And the data only remains in the actual source databases.

Issues in Data Integration:

- There are no of issues to consider during data integration: Schema Integration, Redundancy, Detection and resolution of data value conflicts. These are explained in brief as following below.

1. Schema Integration:

- Integrate metadata from different sources.
- The real world entities from multiple source be matched referred to as the entity identification problem.
- For example, How can the data analyst and computer be sure that customer id in one data base and customer number in another reference to the same attribute.

2. Redundancy:

- An attribute may be redundant if it can be derived or obtaining from another attribute or set of attribute.
- Inconsistencies in attribute can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis.

3. Detection and resolution of data value conflicts:

- This is the third important issues in data integration.
- Attribute values from another different sources may differ for the same real world entity.
- An attribute in one system may be recorded at a lower level abstraction then the “same” attribute in another.

Data Transformation Defined

- Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system.
- Data transformation is a component of most data integration and data management tasks, such as data wrangling and data warehousing.
- One step in the ELT/ETL process, data transformation may be described as either “simple” or “complex,” depending on the kinds of changes that must occur to the data before it is delivered to its target destination.
- The data transformation process can be automated, handled manually, or completed using a combination of the two.
- An ever-increasing number of programs, applications, and devices continually produce massive volumes of data. And with so much disparate data streaming in from a variety of sources, data compatibility is always at risk.
- The data transformation process comes in: it allows companies and organizations to convert data from any source into a format that can be integrated, stored, analyzed, and ultimately mined for actionable business intelligence.

Benefits of Data Transformation

The challenge here is to make sure that all the data that's being collected can be used. By using a data transformation process, companies are able to reap massive benefits from their data, including:

•**Getting maximum value from data**: Forrester reports that between 60 percent and 73 percent of all data is never analyzed for business intelligence. Data transformation tools allow companies to standardize data to improve accessibility and usability.

•**Managing data more effectively**: With data being generated from an increasing number of sources, inconsistencies in metadata can make it a challenge to organize and understand data. Data transformation refines metadata to make it easier to organize and understand what's in your data set.

•**Performing faster queries**: Transformed data is standardized and stored in a source location, where it can be quickly and easily retrieved.

•**Enhancing data quality**: Data quality is becoming a major concern for organizations due to the risks and costs of using bad data to obtain business intelligence. The process of transforming data can reduce or eliminate quality issues like inconsistencies and missing values.

PT_ID	A	B	C	D
1			10	
2	10	20		
3		90%		20%

Structure transformation

PT_ID	TEST	SCORE
1	C	10
2	B	20
2	A	10
3	D	20%
3	B	90%

I

II

Value normalization standardization and validation

PT_ID	TEST	SCORE	SCORE_PCT	ASMT	METHOD
1	C	10		Positive	Panel-A
2	B	20		Negative	Panel-A
2	A	10		Positive	Panel-A
3	D		20	Negative	Panel-B
3	B		90	Positive	Panel-B

Data reduction

- **Data reduction** is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form.
- The basic concept is the **reduction** of multitudinous amounts of **data** down to the meaningful parts.
- Data reduction is the process of reducing the amount of **capacity** required to store data.
- Data reduction can increase storage efficiency and reduce costs. Storage vendors will often describe storage **capacity** in terms of raw **capacity** and effective **capacity**, which refers to data after the reduction.

Data Reduction In Data Mining

- A database or date warehouse may store terabytes of data.So it may take very long to perform data analysis and mining on such huge amounts of data.
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contain critical information.

Data Reduction Strategies:-

1. Data Cube Aggregation

- Aggregation operations are applied to the data in the construction of a data cube.

2. Dimensionality Reduction

- In dimensionality reduction redundant attributes are detected and removed which reduce the data set size.

3. Data Compression

- Encoding mechanisms are used to reduce the data set size.

4. Numerosity Reduction

- In numerosity reduction where the data are replaced or estimated by alternative.

5. Discretization and concept hierarchy generation

- Where raw data values for attributes are replaced by ranges or higher conceptual levels.

RollNo	Name	Mobile Number	Mobile Network
T4Tutorials1	Sameed	+92 302 XX XXX XX	Mobilink
T4Tutorials1	Ali	+92 333 XX XXX XX	Ufone

Figure 1: Before Dimension reduction

If we know Mobile Number, then we can know the Mobile Network. So we need to reduce the one dimension.

RollNo	Name	Mobile Number
T4Tutorials1	Sameed	+92 302 XX XXX XX
T4Tutorials1	Ali	+92 333 XX XXX XX

Figure 2: After Dimension reduction

Unit –III : Data Mining Query Language

Generalization and Summarization

- Data and objects in databases contain detailed information at the primitive concept level.
- For example, the item relation in a sales database may contain attributes describing low-level item information such as item_ID, name, brand, category, supplier, place_made and price.
- It is useful to be able to summarize a large set of data and present it at a high conceptual level.
- For example, summarizing a large set of items relating to Christmas season sales provides a general description of such data, which can be very helpful for sales and marketing managers.
- This requires an important functionality called data generalization.

Generalization and Summarization

- A process that abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.
- Data Generalization is a summarization of general features of objects in a target class and produces what is called characteristic rules.
- The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.
- For example, one may want to characterize the "OurVideoStore" customers who regularly rent more than 30 movies a year.
- With concept hierarchies on the attributes describing the target class, the attribute-oriented induction method can be used, for example, to carry out data summarization.

Generalization and Summarization

- Note that with a data cube containing a summarization of data, simple OLAP operations fit the purpose of data characterization.
- **Approaches:**
    1. Data cube approach(OLAP approach).
    2. Attribute-oriented induction approach.

Presentation Of Generalized Results

Generalized Relation:

Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.

Cross-Tabulation:

Mapping results into cross-tabulation form (similar to contingency tables).

Generalization and Summarization

Visualization Techniques:

- ❖ Pie charts, bar charts, curves, cubes, and other visual forms.

Quantitative characteristic rules:

- ❖ Mapping generalized results in characteristic rules with quantitative information associated with it.

Summarization

- Data Summarization is a simple term for a short conclusion of a big theory or a paragraph.
- This is something where you write the code and, in the end, you declare the final result in the form of summarizing data.
- Data summarization has the great importance in the data mining.
- As nowadays a lot of programmers and developers work on big data theory. Earlier, you used to face difficulties to declare the result, but now there are so many relevant tools in the market where you can use in the programming or wherever you want in your data.
- **Summarization** is a key **data mining** concept which in- volves techniques for finding a compact description of a dataset.
- Simple **summarization** methods such as tabulating the mean and standard deviations are often applied for exploratory **data** analysis, **data** visualization and automated report generation.

Summarization

Data Summarization

You can make the summary of the data in excel with ease and in a less time. There are so many ways to mining the data in excel, but I tell you the very simple formula for summarizing the data. First thing, we need a table as follows,

Data Summarization with SUMIF()

Now, look in the table name Register which contains six columns such as Invoice Date, Customer, Type, Country, Amount, and Status. We have already inserted random data for our ease. Now, I want total amount of 'Paid' and 'Future' in the table. So, let's understand with an example of SUMIF formula as follows,

Total Amount of Invoices?

Paid	
Future	

Register

Invoice Date	Customer	Type	Country	Amount	Status
04-05-2018	Dominoz	Manufacturer	India	54000	Paid
08-05-2018	McDonald	Retailer	USA	14560	Due
15-06-2018	KFC	Distributor	Canada	13500	Future
07-05-2018	Subway	Manufacturer	USA	69500	Due
06-02-2018	KFC	Distributor	India	12500	Paid
07-02-2018	Dominoz	Retailer	Canada	15200	Future
28-02-2018	KFC	Distributor	Canada	56000	Due
24-06-2018	Subway	Manufacturer	USA	25200	Paid
25-06-2018	Dominoz	Distributor	India	32000	Future
26-06-2018	KFC	Retailer	Canada	17100	Paid
27-06-2018	Subway	Manufacturer	India	26800	Paid
29-12-2018	KFC	Distributor	Canada	12650	Due
30-11-2018	Dominoz	Manufacturer	USA	12200	Paid
19-08-2018	Subway	Distributor	India	23003	Paid
20-08-2018	Subway	Retailer	Canada	12400	Future
14-07-2018	Dominoz	Manufacturer	USA	51000	Due

Total Amount of Invoices?

Paid	$ 1,70,803.00
Future	$ 73,100.00

Register

Invoice Date	Customer	Type	Country	Amount	Status
04-05-2018	Dominoz	Manufacturer	India	54000	Paid
08-05-2018	McDonald	Retailer	USA	14560	Due
15-06-2018	KFC	Distributor	Canada	13500	Future
07-05-2018	Subway	Manufacturer	USA	69500	Due
06-02-2018	KFC	Distributor	India	12500	Paid
07-02-2018	Dominoz	Retailer	Canada	15200	Future
28-02-2018	KFC	Distributor	Canada	56000	Due
24-06-2018	Subway	Manufacturer	USA	25200	Paid
25-06-2018	Dominoz	Distributor	India	32000	Future
26-06-2018	KFC	Retailer	Canada	17100	Paid
27-06-2018	Subway	Manufacturer	India	26800	Paid
29-12-2018	KFC	Distributor	Canada	12650	Due
30-11-2018	Dominoz	Manufacturer	USA	12200	Paid
19-08-2018	Subway	Distributor	India	23003	Paid
20-08-2018	Subway	Retailer	Canada	12400	Future
14-07-2018	Dominoz	Manufacturer	USA	51000	Due

ASSOCIATION MINING

- Proposed by Agrawal et al in 1993.
- **Association rule mining** is a procedure which is meant to find frequent patterns, correlations, associations, or causal
- structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- It is an important data mining model studied extensively by the database and data mining community.

Applications in Association Mining

- **Market Basket Analysis:**

  Given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.

- **Telecommunication**

  Each customer is a transaction containing the set of phone calls

- **Credit Cards/ Banking Services**

  Each card/account is a transaction containing the set of customer's payments

- **Medical Treatments**

  Each patient is represented as a transaction containing the ordered set of diseases

MARKET BASKET ANALYSIS

- **INPUT**: list of purchases by purchaser
  - do not have names
- ***identify purchase patterns***
- what items are purchased sequentially
  - obvious: house-furniture; car-tires
- what items tend to be purchased by season
- Categorize customer purchase behavior
- identify ***actionable*** information
  - purchase profiles
  - profitability of each purchase profile
  - use for marketing
    - layout or catalogs
    - select products for promotion
    - space allocation, product placement

POSSIBLE MARKET BASKETS

Customer 1: diapers, baby lotion, grapefruit juice, baby food, milk

Customer 2: soda, potato chips, milk

Customer 3: soup, beer, milk, ice cream

Customer 4: soda, coffee, milk, bread

Customer 5: beer, potato chips

CO-OCCURRENCE TABLE

Customer 1: diapers, baby lotion, grapefruit juice, baby food, milk

Customer 2: soda, potato chips, milk

Customer 3: soup, beer, milk, ice cream

Customer 4: soda, coffee, milk, bread

Customer 5: beer, potato chips

	Beer	Pot. Chips	Milk	Diap.	Soda
Beer	3	2	1	0	0
Pot. Chips	2	3	1	0	1
Milk	1	2	4	1	2
Diapers	0	0	1	1	0
Soda	0	1	2	0	2

beer & potato chips - makes sense

milk & soda - probably noise

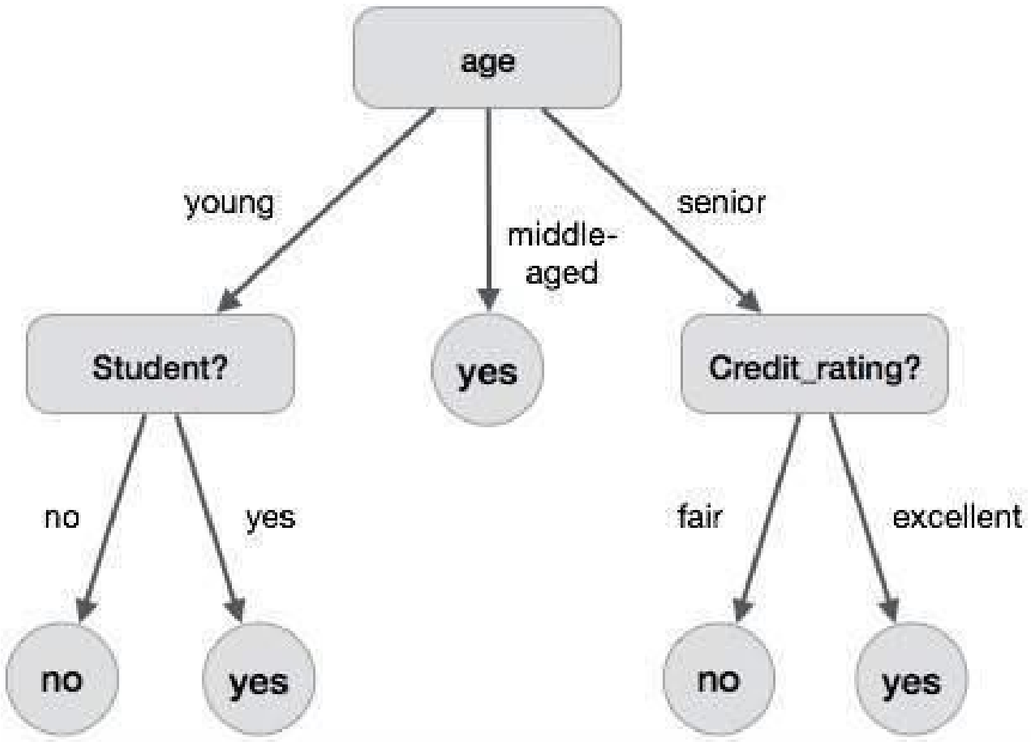
Unit - IV

Decision Tree Induction

- A decision tree is a structure that includes a root node, branches, and leaf nodes.
- Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.
- The topmost node in the tree is the root node.
- Each internal node represents a test on an attribute. Each leaf node represents a class.
- The benefits of having a decision tree are as follows −
    - It does not require any domain knowledge.
    - It is easy to comprehend.
    - The learning and classification steps of a decision tree are simple and fast.

Decision Tree Induction

❖ The following decision tree is for the concept buy computer that indicates whether a customer at a company is likely to buy a computer or not.



Decision Tree Induction Algorithm

- A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser).
- Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach.
- In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner.

Tree Pruning

- Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Decision Tree Induction Algorithm

❖ Tree Pruning Approaches

There are two approaches to prune a tree.

1. Pre-pruning – The tree is pruned by halting its construction early.
2. Post-pruning - This approach removes a sub-tree from a fully grown tree.

❖ Cost Complexity

❖ The cost complexity is measured by the following two parameters –

❖ Number of leaves in the tree, and

❖ Error rate of the tree.

Unit – IV : Classification

- Classification models predict categorical class labels; and prediction models predict continuous valued functions.
- There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follow.
  - Classification
  - Prediction
- For example, we can build a classification model to categorize bank loan applications as either safe or risky.
- Prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

What is classification (Tom Mitchill)

- It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.
- Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.
- **Example**: Before starting any Project, we need to check it's feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two-step process such as :

Unit – IV : Classification

How Does Classification Works?

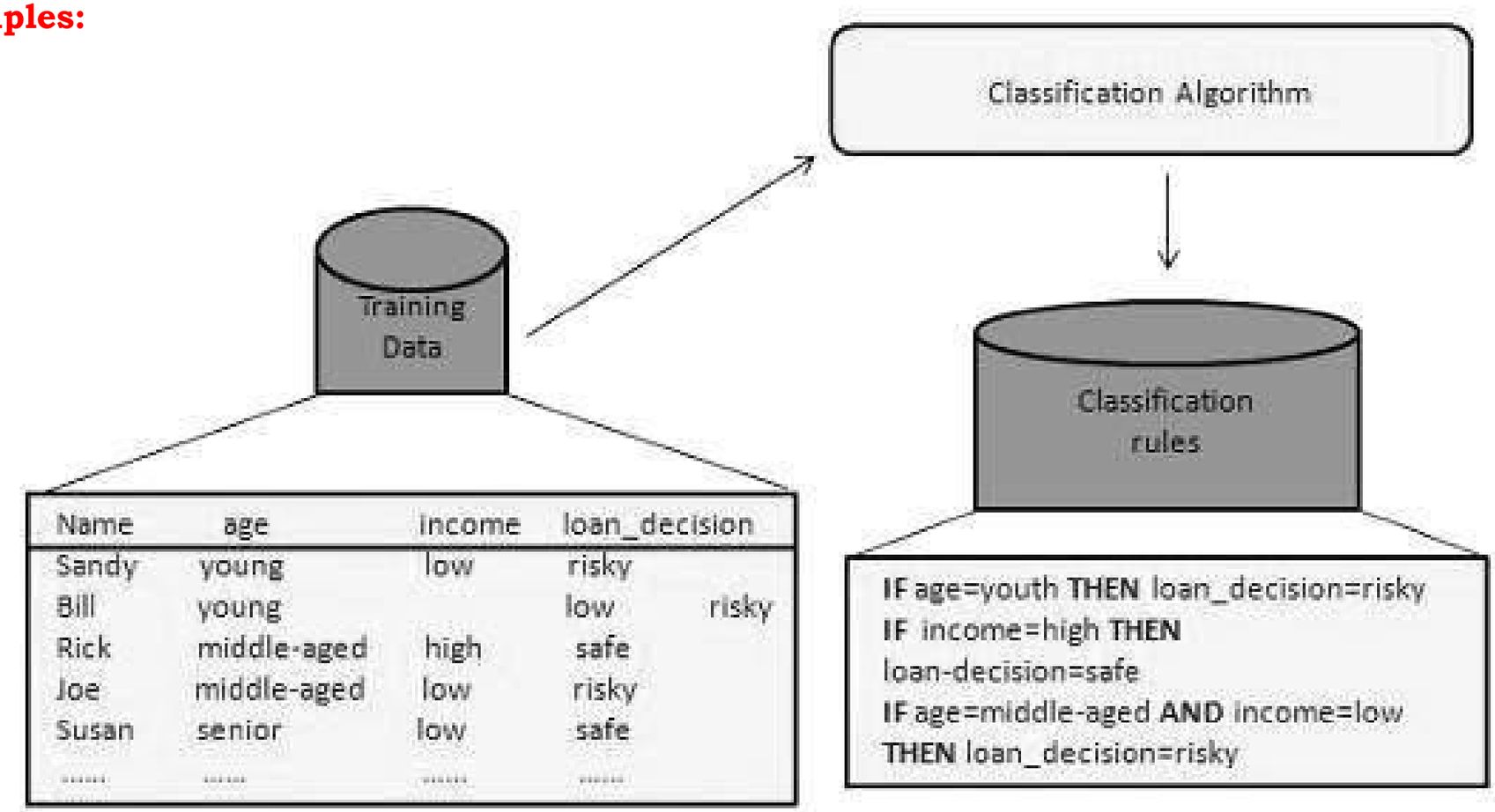
❖ With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps

1. Building the Classifier or Model
2. Using Classifier for Classification

Building the Classifier or Model

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

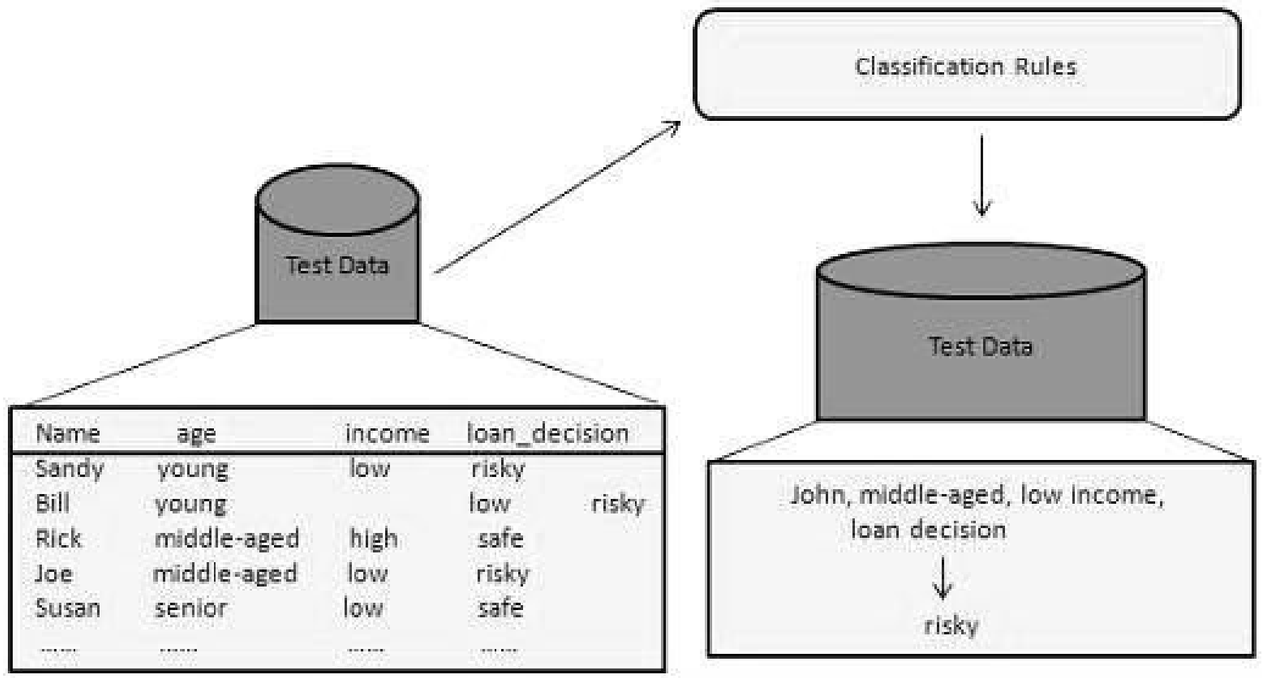
Examples:


Classification Algorithm
Training Data
Classification rules
Name age income loan_decision
Sandy young low risky
Bill young low risky
Rick middle-aged high safe
Joe middle-aged low risky
Susan senior low safe
IF age=youth THEN loan_decision=risky
IF income=high THEN loan-decision=safe
IF age=middle-aged AND income=low THEN loan_decision=risky

Unit – IV : Classification

Using Classifier for Classification

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



Unit – IV : Data Mining - Bayesian Classification

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities −

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

P(H/X)= P(X/H)P(H) / P(X)

Unit – IV : Data Mining - Bayesian Classification

Bayesian Belief Network

Bayesian Belief Networks specify joint conditional probability distributions. They are also known as Acceptance Networks, Bayesian Networks, or Probabilistic Networks.

•A Belief Network allows class conditional independencies to be defined between subsets of variables.

•It provides a graphical model of causal relationship on which learning can be performed.

•We can use a trained Bayesian Network for classification.

There are two components that define a Bayesian Belief Network –

•Directed acyclic graph

•A set of conditional probability tables

Directed Acyclic Graph

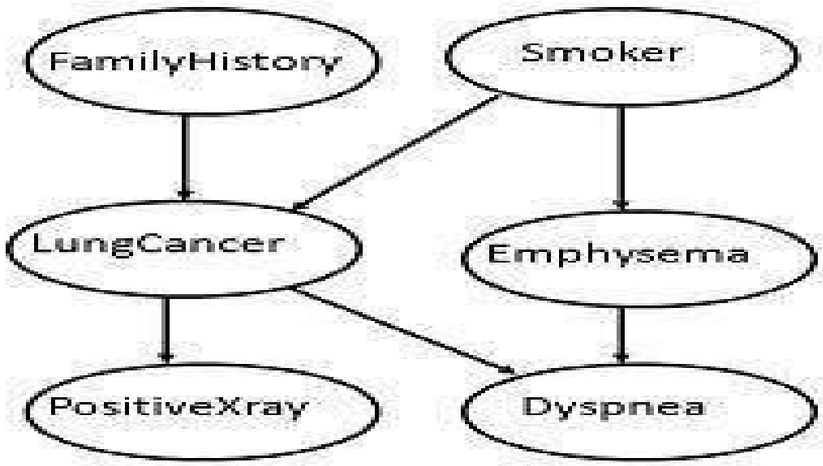
•Each node in a directed acyclic graph represents a random variable.

•These variable may be discrete or continuous valued.

•These variables may correspond to the actual attribute given in the data.

Directed Acyclic Graph Representation

The following diagram shows a directed acyclic graph for six Boolean variables.



- The arc in the diagram allows representation of causal knowledge.
- For example, lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker.
- It is worth noting that the variable PositiveXray is independent of whether the patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has lung cancer.

Conditional Probability Table

The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH), and Smoker (S) is as follows –

	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

Unit – IV : Prediction

- Prediction in data mining is to identify data points purely on the description of another related data value.
- It is not necessarily related to future events but the used variables are unknown.
- Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.
- For example, prediction models in data mining are used by a marketing manager who predict that how much amount a particular customer will spend during a sale, so that upcoming sale amount can be planned accordingly.
- The prediction in data mining is known as Numeric Prediction. Generally regression analysis is used for prediction.

Unit – IV : Prediction

- Predicting the identity of one thing based purely on the description of another, related thing.
- Not necessarily future events, just unknowns
- Based on the relationship between a thing that you can know and a thing you need to predict

Unit – IV : Prediction

Usual Examples

- Predicting levels of sales that will result from a price change or advert.
- Predicting whether or not it will rain based on current humidity
- Predicting the colour of a pottery glaze based on a mixture of base pigments
- Predicting how far up the charts a single will go Predicting how much revenue a book of debt will bring

Unit – IV : Prediction

- Prediction in data mining is to identify data points purely on the description of another related data value.
- It is not necessarily related to future events but the used variables are unknown.
- Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.
- For example, prediction models in data mining are used by a marketing manager who predict that how much amount a particular customer will spend during a sale, so that upcoming sale amount can be planned accordingly.
- The prediction in data mining is known as Numeric Prediction. Generally regression analysis is used for prediction.

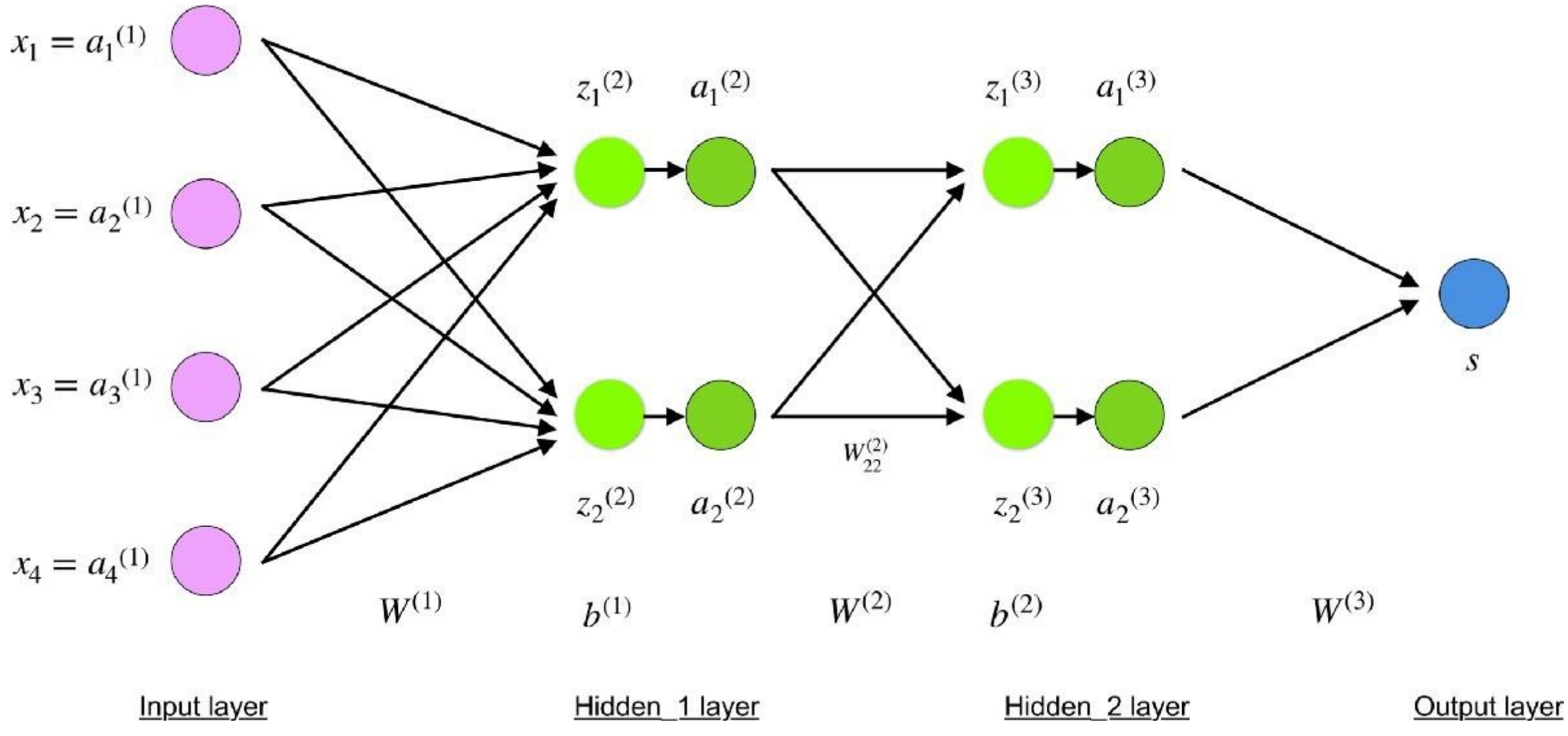
Weather conditions for June 3, 2014

Unit – IV : Backpropagation

- Backpropagation algorithm is probably the most fundamental building block in a neural network. It was first introduced in 1960s and almost 30 years later (1989) popularized by Rumelhart, Hinton and Williams in a paper called “Learning representations by back-propagating errors”.
- Backpropagation (backward propagation) is an important mathematical tool for improving the accuracy of predictions in data mining and machine learning.
- Essentially, backpropagation is an algorithm used to calculate derivatives quickly.
- Artificial neural networks use backpropagation as a learning algorithm to compute a gradient descent with respect to weights.
- Desired outputs are compared to achieved system outputs, and then the systems are tuned by adjusting connection weights to narrow the difference between the two as much as possible.
- The algorithm gets its name because the weights are updated backwards, from output towards input.

Unit – IV : Backpropagation

- The algorithm is used to effectively train a neural network through a method called chain rule.
- In simple terms, after each forward pass through a network, backpropagation performs a backward pass while adjusting the model's parameters (weights and biases).
- I would like to go over the mathematical process of training and optimizing a simple 4-layer neural network.
- I believe this would help the reader understand how backpropagation works as well as realize its importance.
- **Define the neural network model**
- The 4-layer neural network consists of 4 neurons for the input layer, 4 neurons for the hidden layers and 1 neuron for the output layer.



Simple 4-layer neural network illustration

Unit – IV : Backpropagation

Why Need Backpropagation?

- This is a mechanism used to train the neural network relating to the particular dataset.

Some of the advantages of Backpropagation are,

- It is simple, fast and easy to program
- Only numbers of the input are tuned and not any other parameter
- No need to have prior knowledge about the network
- It is flexible
- A standard approach and works efficiently
- It does not require the user to learn special functions

Disadvantages of Backpropagation

- Backpropagation possibly be sensitive to noisy data and irregularity
- The performance of this is highly reliant on the input data
- Needs excessive time for training
- The need for a matrix-based method for backpropagation instead of mini-batch

Applications of Backpropagation

- The neural network is trained to enunciate each letter of a word and a sentence
- It is used in the field of speech recognition
- It is used in the field of character and face recognition

https://www.elprocus.com/what-is-backpropagation-neural-network-types-and-its-applications/

Unit – IV : Clustering Analysis

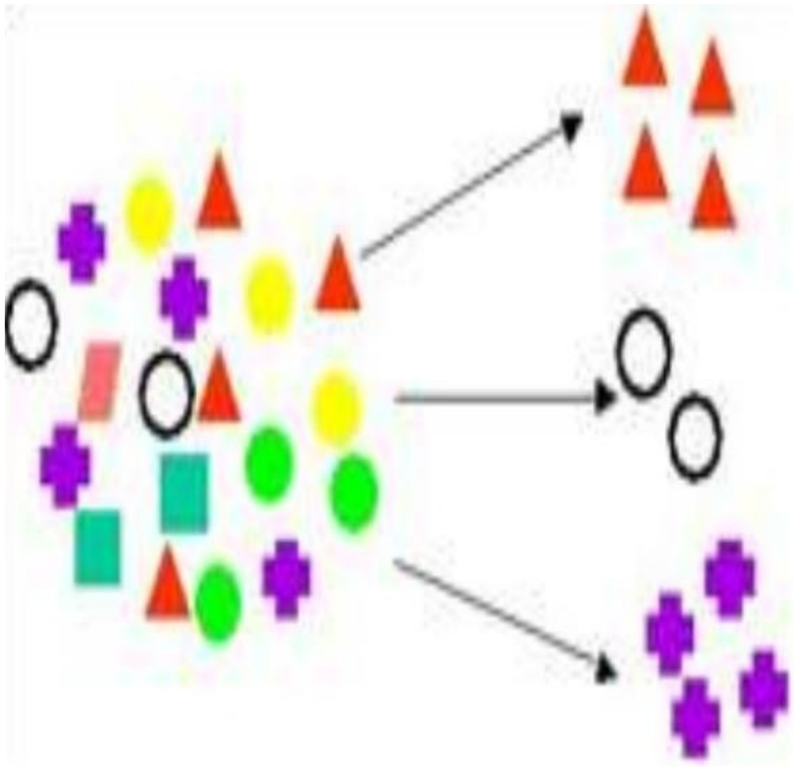
What is Clustering in Data Mining?

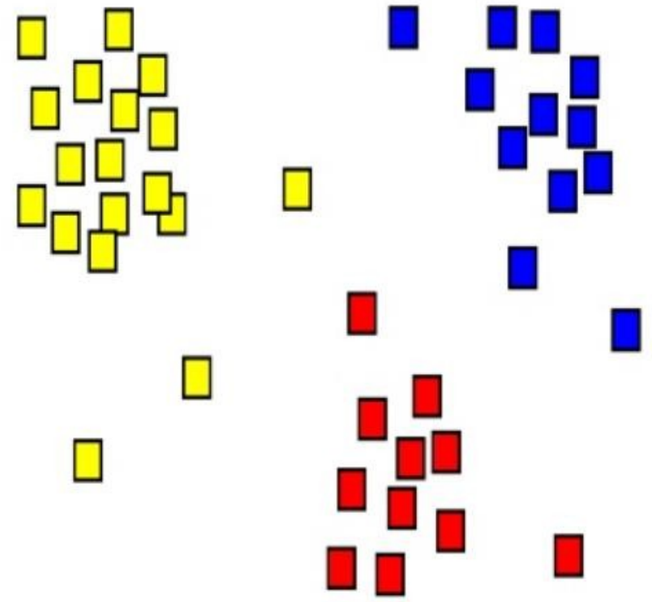
- Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.
- In clustering, a group of different data objects is classified as similar objects.
- One group means a cluster of data.
- Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data.
- After the classification of data into various groups, a label is assigned to the group.
- It helps in adapting to the changes by doing the classification.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Examples for Clustering





https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Applications of Cluster Analysis:

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. Characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities.
- Clustering also helps in identification of areas of similar land use in an earth observation database.
- It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Clustering Methods:

Clustering methods can be classified into the following categories.

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Hierarchical Methods

- A Hierarchical clustering method works via grouping data into a tree of clusters.
- Hierarchical clustering begins by treating every data points as a separate cluster.
- Then, it repeatedly executes the subsequent steps:
    1. Identify the 2 clusters which can be closest together.
    2. Merge the 2 maximum comparable clusters.
- We need to continue these steps until all the clusters are merged together.
- In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters.
- A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).
- The basic method to generate hierarchical clustering are:
    1. Agglomerative (திரட்டுதல்) Approach
    2. Divisive (பிளவுபடுத்தும்) Approach

Agglomerative Approach:

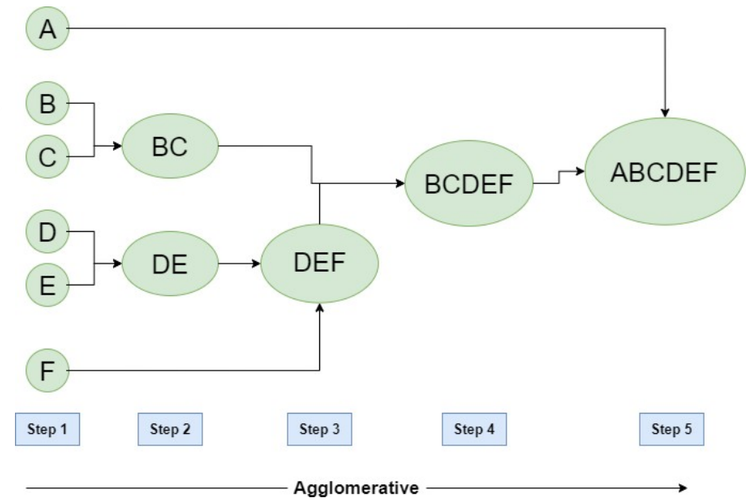
- ❖ Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method).
- ❖ At first every dataset is considered as individual entity or cluster.
- ❖ At every iteration, the clusters merge with different clusters until one cluster is formed.

Algorithm for Agglomerative Hierarchical Clustering is:

1. Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
2. Consider every data point as a individual cluster
3. Merge the clusters which are highly similar or close to each other.
4. Recalculate the proximity matrix for each cluster
5. Repeat Step 3 and 4 until only a single cluster remains.

Note:

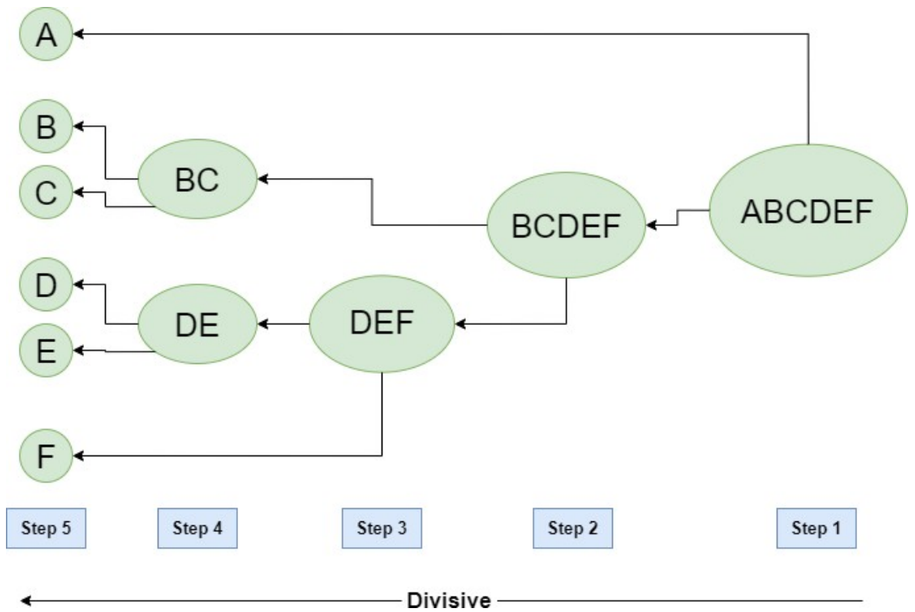
This is just a demonstration of how the actual algorithm works no calculation has been performed below all the proximity among the clusters are assumed.



https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/

Divisive Approach:

- We can say that the Divisive Hierarchical clustering is precisely the opposite of the Agglomerative Hierarchical clustering.
- In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable.
- In the end, we are left with N clusters.



https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/

Unit – IV : Clustering Analysis

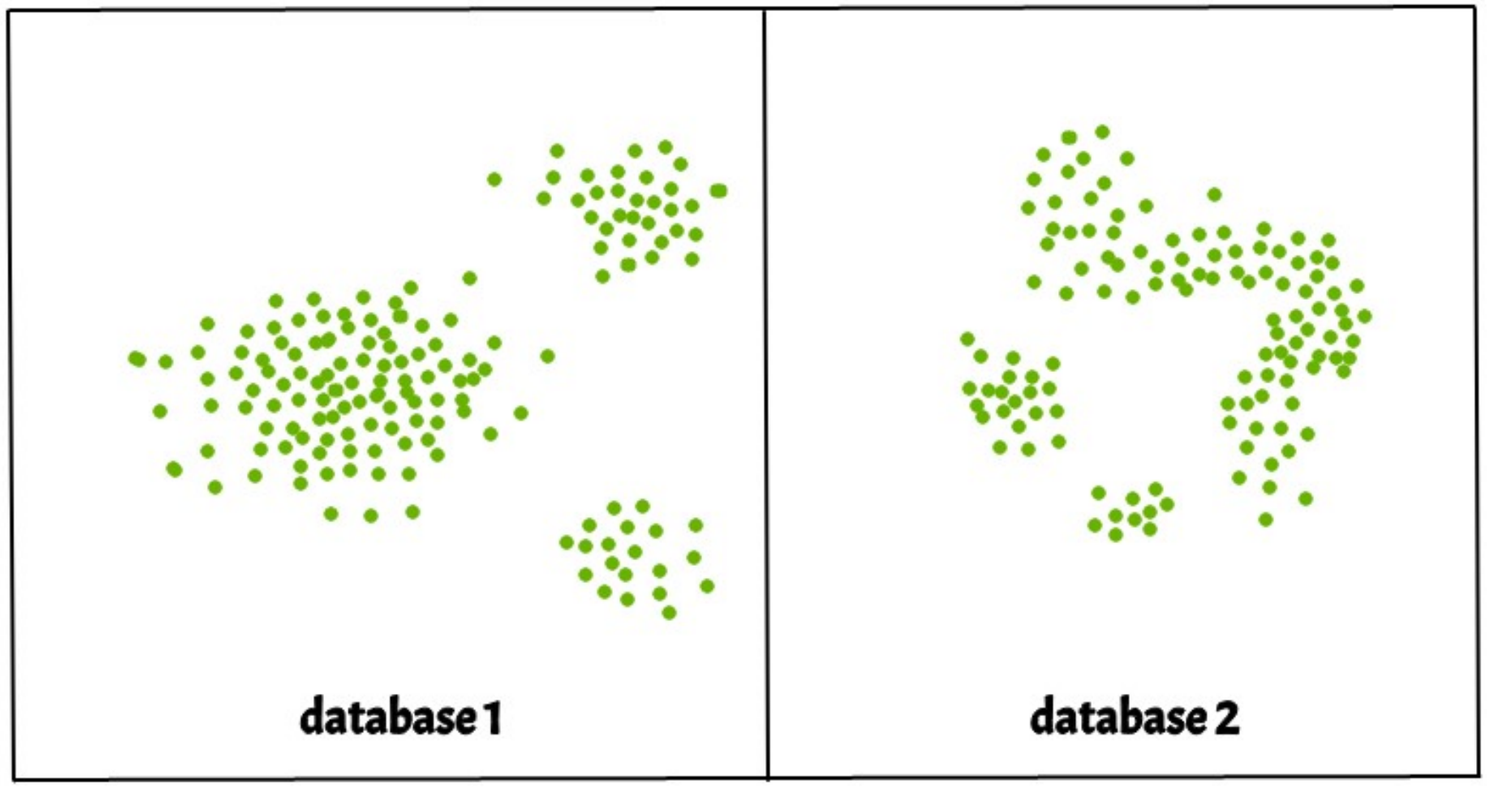
Density-based Method:

- This method is based on the opinion of density.
- Fundamentally, all clustering methods use the same approach i.e. first we calculate similarities and then we use it to cluster the data points into groups or batches.
- Here we will focus on Density-based spatial clustering of applications with noise (DBSCAN) clustering method.
- Clusters are dense regions in the data space, separated by regions of the lower density of points.
- The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”.
- The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

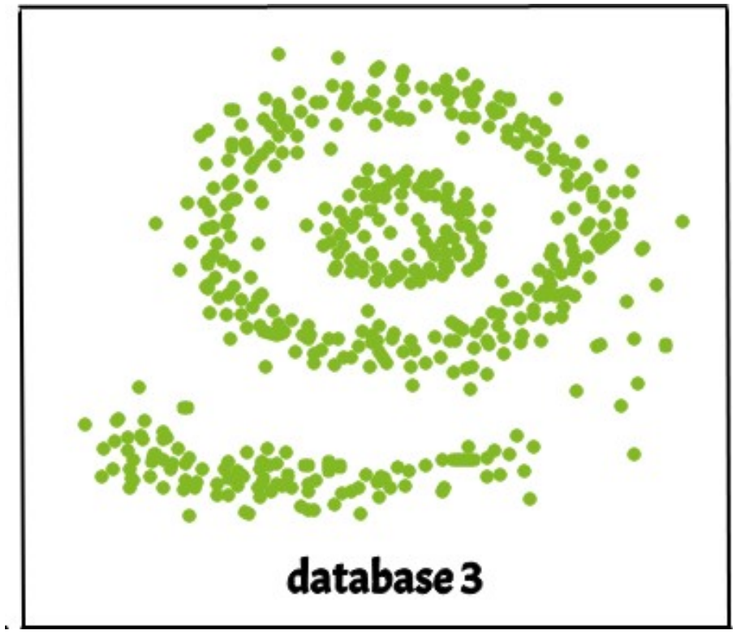
Density-based Method:



https://www.upgrad.com/blog/cluster-analysis-data-mining/

Why DBSCAN ?

- Partitioning methods (K-means, PAM clustering) and hierarchical clustering work for finding spherical-shaped clusters or convex clusters.
- In other words, they are suitable only for compact and well-separated clusters.
- Moreover, they are also severely affected by the presence of noise and outliers in the data.
- Real life data may contain irregularities, like.
  i) Clusters can be of arbitrary shape such as those shown in the figure below.
  ii) Data may contain noise.



https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Grid-Based Clustering Algorithms:

- Density-based and/or grid-based approaches are popular for mining clusters in a large multidimensional space wherein clusters are regarded as denser regions than their surroundings.
- The computational complexity of most clustering algorithms is at least linearly proportional to the size of the data set.
- The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Grid-Based Clustering Algorithms:

- The grid-based clustering approach differs from the conventional clustering algorithms.
- It is concerned not with the data points but with the value space that surrounds the data points.
- In general, a typical grid-based clustering algorithm consists of the following five basic steps (Grabusts and Borisov, 2002):

1. Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
2. Calculating the cell density for each cell.
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

Unit – IV : Clustering Analysis

Grid-based Method

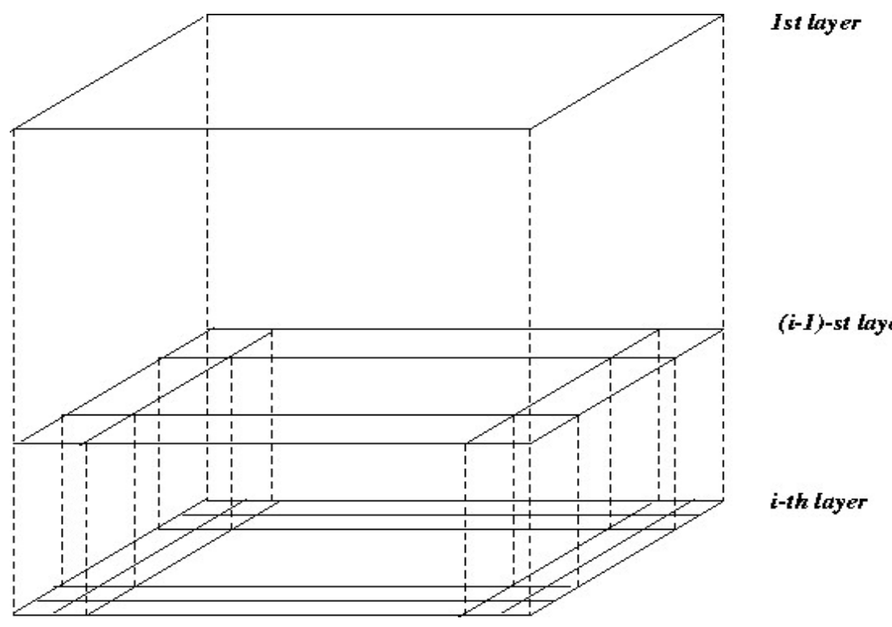
- ❖ In this, the objects together form a grid.
- ❖ The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

https://www.upgrad.com/blog/cluster-analysis-data-mining/

STING: A Statistical Information Grid Approach



Grid-Based Clustering Methods

- Using multi-resolution grid data structure
- Clustering complexity depends on the number of populated grid cells and not on the number of objects in the dataset
- Several interesting methods (in addition to the basic grid-based algorithm)
  - STING (a STatistical INformation Grid approach)
  - CLIQUE

Unit – IV : Outlier Analysis

- An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error.
- The analysis of outlier data is referred to as outlier analysis or outlier mining.

Why outlier analysis?

- Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such case.

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit – IV : Outlier Analysis

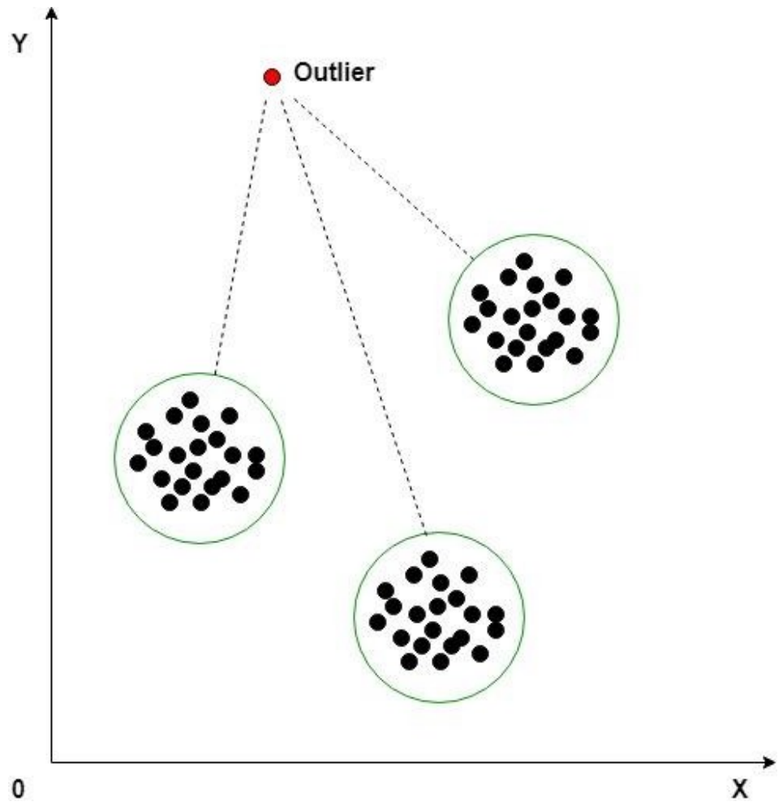


Fig. Outlier analysis

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit – IV : Outlier Analysis

Detecting Outlier:

- ❖ Clustering based outlier detection using distance to the closest cluster:
- ❖ In the K-Means clustering technique, each cluster has a mean value.
- ❖ Objects belong to the cluster whose mean value is closest to it.
- ❖ In order to identify the Outlier, firstly we need to initialize the threshold value such that any distance of any data point greater than it from its nearest cluster identifies it as an outlier for our purpose.
- ❖ Then we need to find the distance of the test data to each cluster mean.
- ❖ Now, if the distance between the test data and the closest cluster to it is greater than the threshold value then we will classify the test data as an outlier.

https://www.geeksforgeeks.org/machine-learning-outlier/

Algorithms:

1. Calculate the mean of each cluster

2. Initialize the Threshold value

3. Calculate the distance of the test data from each cluster mean

4. Find the nearest cluster to the test data

5. If (Distance > Threshold) then, Outlier

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit - V

Unit – V : Web Mining and Web Content Mining

- **Web Mining** is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Applications of Web Mining:

- Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
- It is used for Web Searching e.g., Google, Yahoo etc
- Web mining is used to predict user behavior.
- Web mining is very useful of a particular Website and e-service

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit – V : Web Mining and Web Content Mining

- Web mining can be broadly divided into three different types of techniques of mining:
  - Web Content Mining,
  - Web Structure Mining,
  - Web Usage Mining.

Web Content Mining:

- Web content mining is the application of extracting useful information from the content of the web documents.
- Web content consist of several types of data – text, image, audio, video etc.
- Content data is the group of facts that a web page is designed.
- It can provide effective and interesting patterns about user needs.
- Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining.

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit – V : Web Mining and Web Content Mining

Web Structure Mining:

- Web structure mining is the application of discovering structure information from the web.
- The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages.
- Structure mining basically shows the structured summary of a particular website.
- It identifies relationship between web pages linked by information or direct link connection.
- To determine the connection between two commercial websites, Web structure mining can be very useful.

https://www.geeksforgeeks.org/machine-learning-outlier/

Unit – V : Web Mining and Web Content Mining

Web Usage Mining:

- Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets.
- And these patterns enable you to understand the user behaviors or something like that.
- In web usage mining, user access data on the web and collect data in form of logs.
- So, Web usage mining is also called log mining.

https://www.geeksforgeeks.org/machine-learning-outlier/