Prof. Vinayashree A.S
Dept of CSE(AI)
MIT Mysore

Module - 1

Machine Learning and Data Essentials

Introduction to Machine Learning

"Computers are able to see, hear and learn. Welcome to the future".

— Dave Waters

Machine learning can enable top management of an organization to extract the knowledge from the data stored in various archives of the business organization to facilitate decision making. Such decisions can be useful for organizations to design new products, improve business processes, and to develop decision support systems.

Need for Machine Learning

Business organizations use huge amount of data for their daily activities.

The full potential of this data was not utilized due to two reasons.

* Data being scattered across different archive systems and organizations not being able to integrate these sources fully.
* The lack of awareness about software tools that could help to unearth the useful information from data.

Machine learning has become so popular because of three reasons.

1. High volume of available data to manage:

Big companies such as Facebook, Twitter and Youtube generate huge amount of data that grows at a phenomenal rate. It is estimated that the data approximately gets doubled every year.

2. The cost of storage has reduced. The hardware cost has also dropped. Therefore, it is easier now to capture, process, store, distribute and transmit the digital information.

3. Third reason for popularity of machine learning is the availability of complex algorithms now.

With the popularity and ready adaption of machine learning by business organizations, it has become a dominant technology trend now.

The Knowledge Pyramid

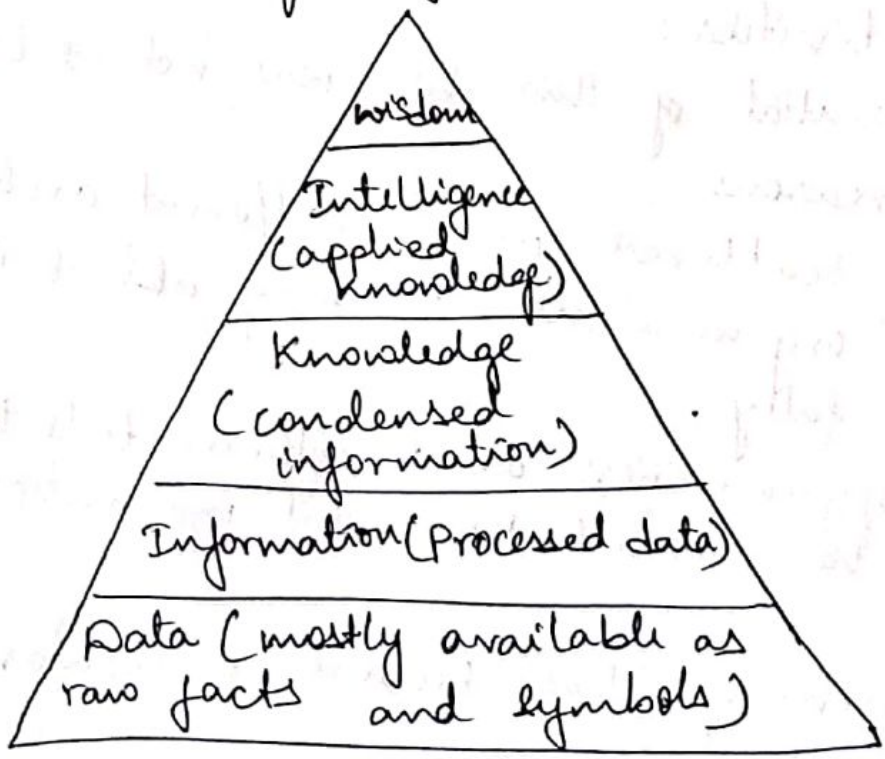


Fig: The Knowledge Pyramid

Data: All facts are data. Data can be numbers or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data with data sources such as flat files, databases, data warehouses in different storage formats.

Information: Processed data is called information. This includes patterns, association or relationships among data.

Ex: Sales data can be analyzed to extract information like which is the fast selling product

Knowledge: Condensed information is called knowledge

Ex: The historical patterns and future trends in the sales data.

Unless knowledge is extracted, data is of no use.

Intelligence: Knowledge is not useful unless it is put into action. Intelligence is the applied knowledge for actions. An actionable form of knowledge is called intelligence.

Wisdom: The ultimate objective of knowledge pyramid is wisdom that represents the maturity of mind that is so, far exibited only by humans

Here comes the need for machine learning. The objective of machine learning is to process these archival data for organizations to take better decisions to design new products, improve the business processes, and to develop effective decision support systems

Machine Learning Explained

Machine learning is an important sub-branch of Artificial Intelligence (AI).

"Machine Learning is the field of study that gives the computer ability to learn without being explicitly programmed."

The system should learn by itself without explicit programming.

It is widely known that to perform a computation one needs to write programs that teach the computers how to do that computation.

In conventional programming, after understanding the problem, a detailed design of the program such as a flowchart or an algorithm needs to be created and coverted into programs using a suitable programming language.

This approach could be difficult for many real-world problems such as puzzles, games, and complex image recognition applications.

Initially, artificial intelligence aims to understand these problems and develop general purpose rules manually. Then, these rules are formulated into logic and implemented in a program to create intelligent systems. This idea of developing intelligent systems is called an expert system.

Ex: MYCIN was designed for medical diagnosis after converting the expert knowledge of many doctors into a system.

Programs still depended on human expertise and hence did not truly exhibit intelligence.

The focus of AI is to develop intelligent systems by using data-driven approach, where data is used as an input to develop intelligent models. The models can then be used to predict new inputs.

Thus, the aim of machine learning is to learn a model or set of rules from the given dataset automatically so that it can predict the unknown data correctly.

As human take decisions based on an experience, computers make models based on extracted patterns in the input data and then use these data-filled models for prediction and to take decisions.

For computers, the learnt model is equivalent to human experience.

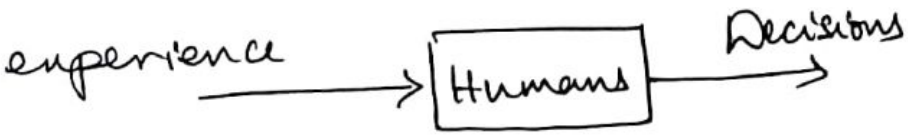


Fig: A learning system for Humans

Fig: A learning system for machine learning

The quality of data determines the quality of experience and therefore the quality of the learning system.

In Statistical learning, the relationship between the input x and output y is modeled as a function in the form $y = f(x)$.

Here, f is the learning function that maps the input x to output y.

In machine learning, this is simply called mapping of input to output.

The learning program summarizes the raw data in a model. A model is an explicit description of patterns within the data in the form of

1. Mathematical equation
2. Relational diagrams like trees / graphs
3. Logical if / else rules, or
4. Groupings called clusters.

In summary, a model can be formula, procedure or representation that can generate data decisions. For example, a model can be helpful to examine whether a given email is spam or not.

Another pioneer of AI, Tom Mitchell's definition of machine learning states that, "A computer program is said to learn from Experience E, with respect to task T and some performance measure P, if its performance on T measured by P improves with experience E."

Ex: The task T could be detecting an object in an image. The machine can gain the knowledge of object using training dataset of thousands of images

This is called experience E. So, the focus is to use this experience E for this task of object detection T. The ability of the system to detect the object is measured by performance measures like precision and recall.

* Models of computer systems are equivalent to human experience. Experience is based on data.
* Humans gain experience by various means. They gain knowledge by rote learning. They observe others and imitate it.
* Humans gain a lot of knowledge from teachers and books.
* We learn many things by trial and error.
* Once the knowledge is gained, when a new problem is encountered, humans search for similar past situations and then formulate the heuristics and use that for prediction.

But, in systems, experience is gathered by these steps.

1. Collection of data.
2. Once data is gathered, abstract concepts are formed out of that data. Abstraction is used to generate concepts. This is equivalent to human idea of objects.
   Ex: we have some idea about how an elephant looks like.
3. Generalization converts the abstraction into an actionable form of intelligence.

A. Heuristics are educated guesses for all tasks. Heuristics normally works! But, occasionally, it may fail too. It is not the fault of heuristics as it is just a 'rule of thumb'.

The course correction is done by taking evaluation measures. Evaluation checks the thoroughness of the models and to-do course correction, if necessary, to generate better formulations.

Machine Learning in Relation to Other fields

Machine Learning and Artificial Intelligence

The aim of AI is to develop intelligent agents. An agent can be a robot or any autonomous systems. The resurgence in AI happened due to development of data driven systems. The aim is to find relations and regularities present in the data.

Machine learning is the subbranch of AI, whose aim is to extract the patterns for prediction. It is a broach field that includes learning from examples and other areas like reinforcement learning. The model can take an unknown instance and generate results.

Deep learning is a subbranch of machine learning. In deep learning, the models are constructed using neural network technology. Neural networks are based on the human neuron models. Many neurons form a network connected with the activation functions that trigger further neurons to perform tasks.

Machine Learning, Data Science, Data Mining and Data Analytics

* Data Science is an 'Umbrella' term that encompasses many fields.
* Machine learning starts with data. Therefore, data science and machine learning are interlinked.
* Machine learning is a brach of data science.

Data Science deals with gathering of data for analysis.

* Data Science is a broad field that includes Big data, Data Mining, Data Analytics and Pattern recognition.

⇒ **Big Data**:

Data Science concerns about collection of data. Big data is a field of data science that deals with data's following characteristics:

1. Volume: Huge amount of data is generated by big companies like Facebook, Twitter, Youtube.
2. Variety: Data is available in variety of forms like images, videos and in different formats.
3. Velocity: It refers to the speed at which the data is generated and processed.

Big data is used by many machine learning algo-rithms for applications such as language translation and image recognition.

⇒ **Data Mining**:

Data mining's original genesis is in the business. Data mining aims to extract the hidden patterns that are present in the data. Machine learning aims to use hidden patterns for prediction

⇒ Data Analytics:

Another branch of data Science is data analytics.

It aims to extract useful knowledge from crude data.

There are different types of analytics. predictive data analytics is used for market predictions. Machine learning is closely related to this brach of analytics and shares almost all algorithms.

⇒ Pattern Recognition:

It is an engineering field. It uses machine learning algorithms to extract the features for pattern analysis and pattern classification.

One can views pattern recognition as a specific application of machine learning.

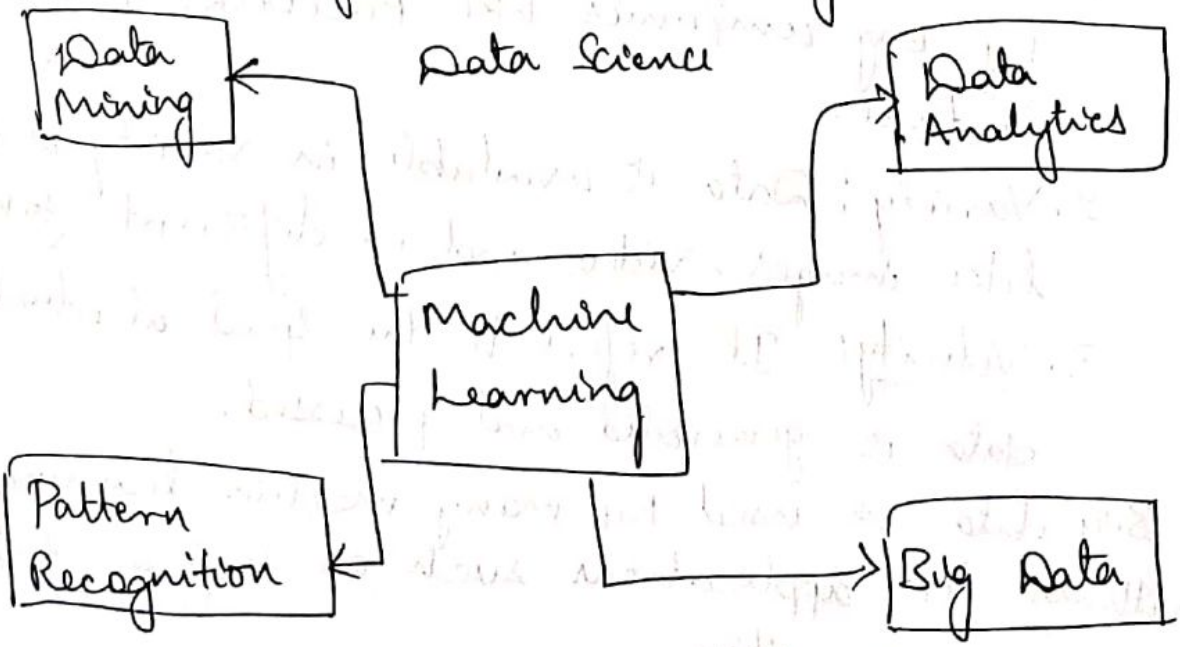


Fig: Relationship of Machine learning with other Major fields

Machine learning and statistics

Statistics is a brach of mathematics that has a solid theoretical foundation regarding statistical learning. Like Machine learning (ML), it can learn from data.

The difference between statistics and ML is that statistical methods look for regularity in data called. patterns. Initially, statistics sets a hypothesis and performs experiments to verify and validate the hypothesis in order to find relationships among data.

statistics is mathematics intensive and models are often complicated equations and involves many assumptions.

Machine learning has less assumptions and requires less statistical knowledge. But, it often requires interaction with various tools to automate the process of learning.

Types of Machine Learning

Learning, like adaption, occurs as the result of interaction of the program with its environment. It can be compared with the interaction between a teacher and a student.

There are four types of machine learning

* Supervised learning
* Unsupervised learning
* Semi-Supervised learning
* Reinforcement learning

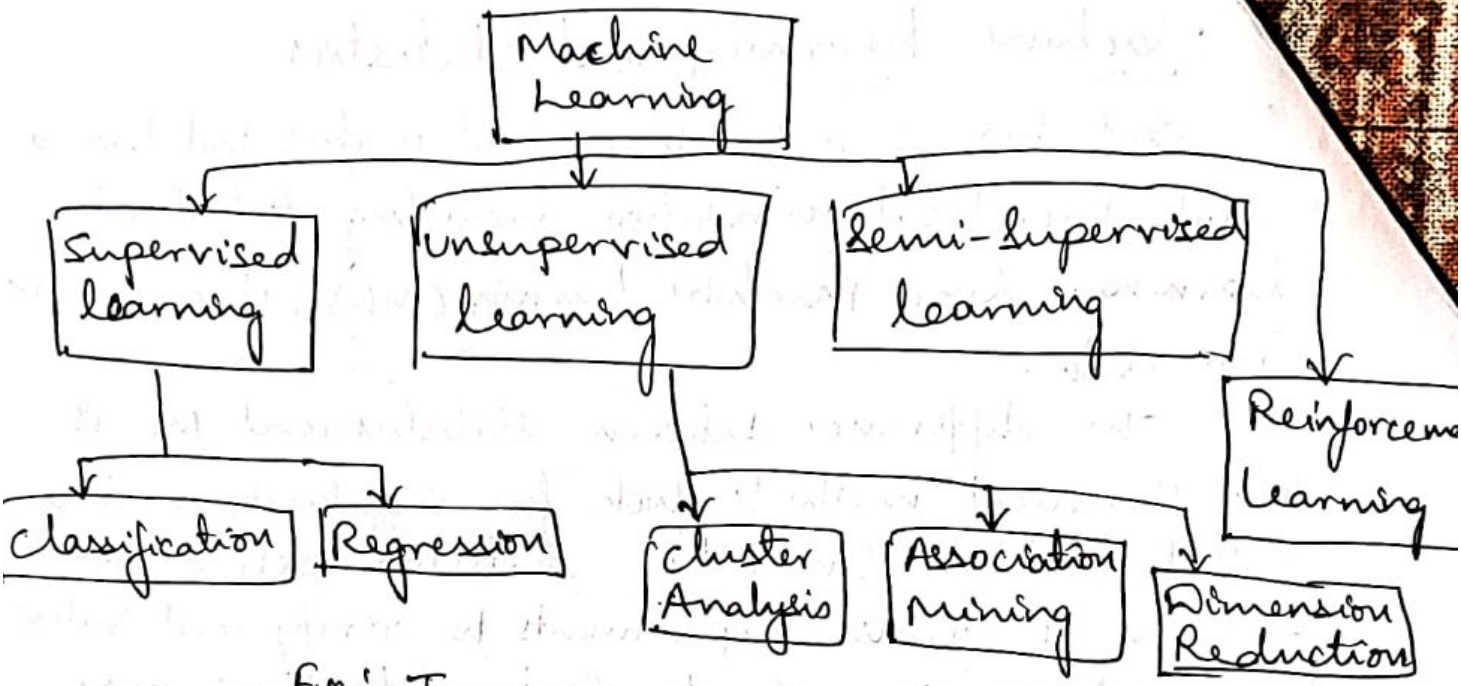


Fig: Types of Machine learning.

Labelled and Unlabelled Data

* Data is a raw fact.
* Normally, data is represented in the form of a table
* Each row of the table represents a data point.
* Features are attributes or characteristics of an object.
* Normally, columns of the table are attributes.
* Label is the feature that we aim to predict.
* There are two types of data – labelled & unlabelled

Labelled Data

In labelled data, there are labels in the dataset.

Ex:

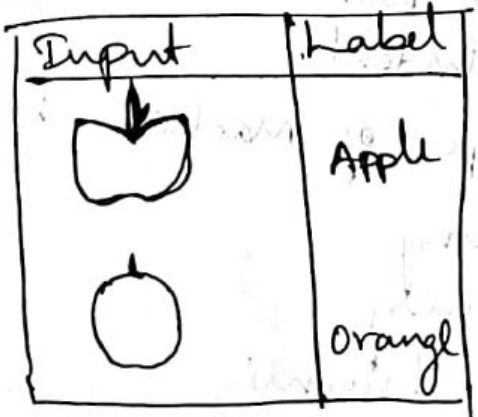


Fig: Labelled dataset

Unlabelled Data

In unlabelled data, there are no labels in the dataset.

Ex!

Fig: Unlabelled Dataset.

* Supervised Learning

* Supervised algorithms use labelled dataset.
* As the name suggests, there is a supervisor or teacher component in supervised learning.
* A supervisor provides labelled data so that the model is constructed and generates test data.
* In supervised learning algorithms, learning takes place in two stages.

In layman terms, during the first stage, the teacher communicates the information to the student that the student is supposed to master. The student receives the information and understands it. During this stage, the teacher has no knowledge of whether the information is grasped by the student.

This leads to the second stage of learning. The teacher then asks the student a set of questions to find out how much information has been grasped by the student. Based on these questions the student is tested, and the teacher informs the student about his assessment. This kind of learning is typically called supervised learning.

* Supervised learning has two methods.
  1. Classification
  2. Regression.

Classification

* Classification is a supervised learning method.
* The input attributes of the classification algorithms are called independent variables.
* The target attribute is called label or dependent variable.
* The relationship between the input and target variable is represented in the form of a structure which is called a classification Model.
* The focus of classification is to predict the 'label' that is in a discrete form.

Ex: A classification algorithm takes a set of labelled data images such as apples and oranges to construct a model that can later be used to classify an unknown test image data.

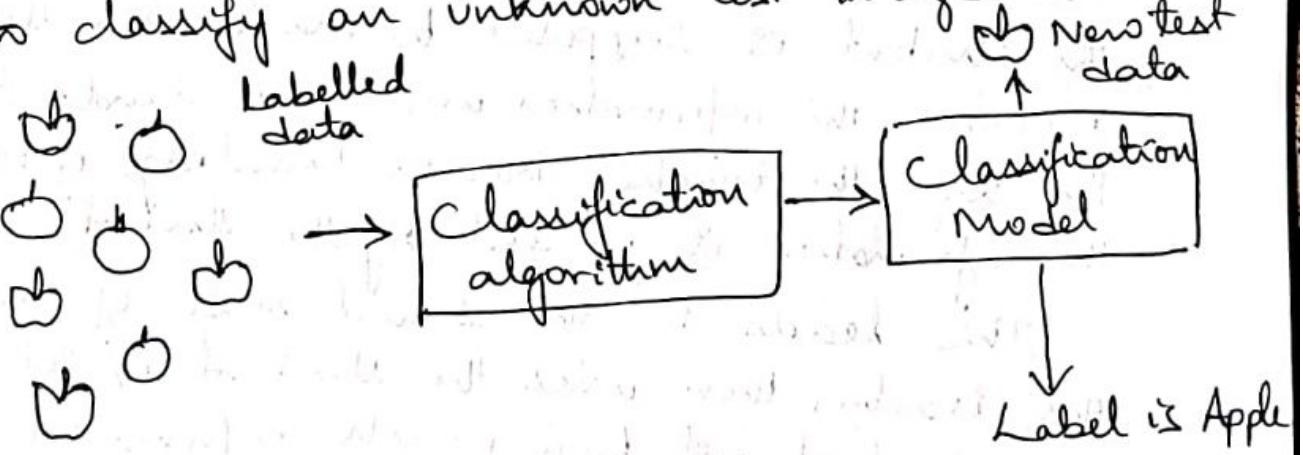


Fig: An example classification system

* In classification, learning takes place in two stages
  - ⇒ Training stage
  - ⇒ Testing stage

Training stage: During training stage, the learning algorithm takes a labelled dataset and starts learning.

Testing stage: The constructed model is tested with test or unknown sample and assigned a label.

This is the classification process.

Classification models can be classified as generative models and discriminative models.

Generative models: It deal with the process of data generation and its distribution.

Ex: Probabilistic models.

Discriminative models: This models do not care about the generation of data. Instead, they simply concentrate on classifying the given data.

Some of the key algorithms of classification are

* Decision Tree
* Random Forest
* Support vector Machines
* Naive Bayes
* Artificial Neural Network and Deep Learning networks like CNN.

Regression Models

Regression Models, unlike classification algorithms predict continuous variables like price. In other words, it is a number.

Ex: For a dataset that represent weeks input x and product sales y. A fitted regression model is as shown in the diagram.

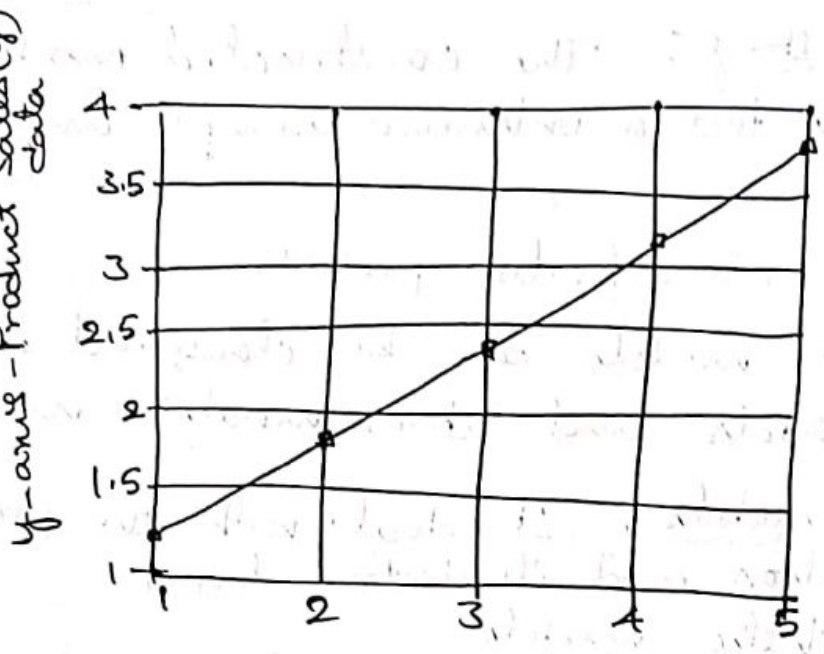


Fig: A Regression Model of the form $y = ax + b$

The regression model takes input x and generates a model in the form of a fitted line of the form $y = f(x)$.

Here, x is the independent variable that may be one or more attributes and y is the dependent variable

The advantage of this model is that prediction for product sales (y) can be made for unknown week data (x). For example, the prediction for unknown 8th week can be made by substituting x as 8 in that regression formula to get y.

The main difference between classification and regression models is that regression models predict continuous variables such as product price, while classification concentrates on assigning labels such as class.

Unsupervised Learning

As the name suggests, there are no supervisor or teacher components. In the absence of a super-visor or teacher, self-instruction is the most common kind of learning process. This process of self-instruction is based on the concept of trial and error.

Here, the program is supplied with objects, but no labels are defined. The algorithm itself observes the examples and recognizes patterns based on the principles of grouping. Grouping is done in ways that similar objects form the same group.

Ex: Cluster analysis, Dimensional Reduction

Cluster Analysis

Cluster analysis is an example of unsupervi-sed learning. It aims to group objects into disjoint clusters or groups.

All the data objects of the partitions are similar in some aspect and vary from the data objects in the other partitions significantly.

Example of clustering processes are — segmenta-tion of a region of interest in an image, detection of abnormal growth in a medical image.

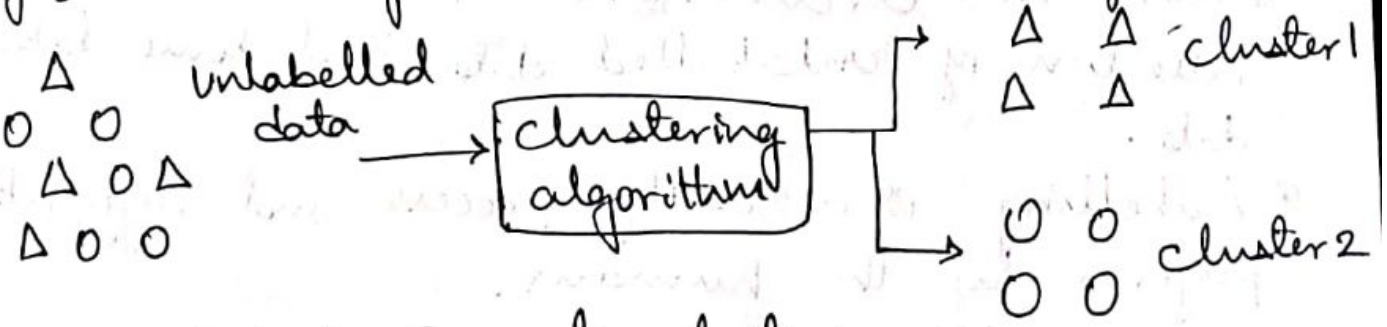


Fig: An Example clustering scheme

Some of the key clustering algorithms are

* K-means algorithm
* Hierarchical algorithms

Dimensionality Reduction

* Dimensionality Reduction algorithms are examples of unsupervised algorithms.
* It takes a higher dimension data as input and outputs the data in lower dimension by taking advantage of the variance of the data.
* It is a task of reducing the dataset with few features without losing the generality.

Differences between Supervised and Unsupervised Learing

Supervised learning	Unsupervised Learning
* There is a supervisor component	* No supervisor component
* Uses Labelled data	* Uses Unlabelled data
* Assign categories or labels	* Performs grouping process such that similar object will be in one cluster.

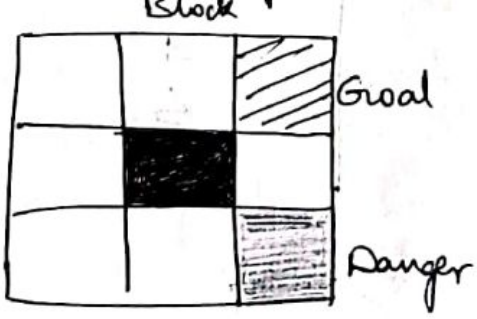
Semi-Supervised Learning

* There are circumstances where the dataset has collection of unlabelled data and some labelled data.
* Labelling is a costly process and difficult to perform by the humans.
* Uses unlabelled data by assigning a Pseudo-labe

* Reinforcement Learning

* Reinforcement learning mimics human beings.
* Like human beings use ears and eyes to perceive the world and take actions, reinforcement learning allows the agent to interact with the environment to get rewards.
* The agent can be human, animal, robot or any independent program.
* The rewards enable the agent to gain experience.
* The agent aims to maximize the reward.
* The reward can be positive or negative (Punishment).
* When the rewards are more, the behavior gets reinforced and learning becomes possible.
* Reinforcement algorithms are reward-based, goal-oriented algorithms

Example: Grid game



In this grid game, the gray tile indicates the danger, black is a block and the tile with diagonal line is the goal.

The aim is to start, say from bottom-left grid; using the actions left, right, top and bottom to reach the goal state.

To solve this sort of problem, there is no data. The agent interacts with the environment to get experience. In the above case, the agent tries to create a model by simulating many paths and finding rewarding paths. This experience helps in constructing a model.

Challenges of Machine Learning

Problems that can be Dealt with Machine Learning

* Machine learning can deal with the 'well-posed' problems where specifications are complete and available. Computers cannot solve 'ill-posed' problems.

Ex:

Input (x_1, x_2)	output (y)
1, 1	1
2, 1	2
3, 1	3
4, 1	4
5, 1	5

$y = x_1 \times x_2$

$y = x_1 \div x_2$

$y = x_1^{x_2}$

There are three functions that fit the data. This means that the problem is ill-posed. To solve this problem, one needs more example

to check the model.

* Huge data: This is a primary requirement of machine learning. Availability of a quality data is a challenge. A quality data means it should be large and should not have data problem -s such as missing data or incorrect data.

* Hogh computation power :- with the availability of Big Data, the computational resource requirement has also increased. Systems with Graphics Processing Unit (GPU) or even Tensor Processing Unit (TPU) are required to execute machine learning algorithms.

* Complexity of the algorithms - It is a challenge for machine learning professionals to design, select, and evaluate optimal algorithms.

* Bias / Variance - Variance is the error of the model. This leads to a problem called bias / variance tradeoff. A model that fits the training data correctly but fails for test data, in general lacks generalization, is called overfitting. The reverse problem is called underfitting where the model fails for training data but has good generalization Overfitting and underfitting are great challenges for machine learning algorithms.

Machine Learning Process

The emerging process model for the data mining solutions for business organizations is CRISP-DM. CRISP-DM stands for Cross Industry Standard Process - Data Mining.

This process involves Six steps

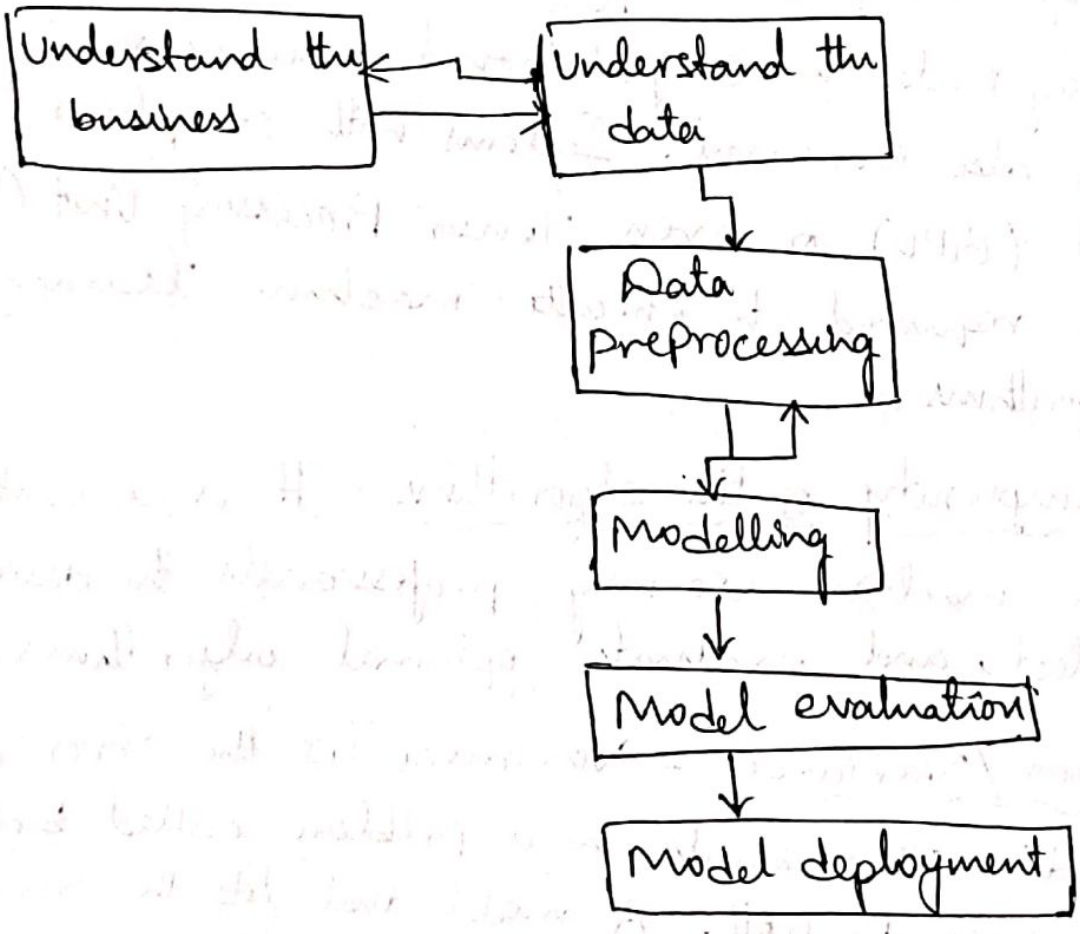


Fig: A Machine Learning / Data Mining Process

1. Understanding the Business:

This step involves understanding the objectives and requirements of the business organization. This step also involves the formulation of the problem statement for the data mining process.

2. Understanding the data:

It involves the steps like data collection, study of the characteristics of the data, formulation of

hypothesis, and matching of patterns to the selected hypothesis.

3. Preparation of data:

This step involves producing the final dataset by cleaning the raw data and preparation of data for the data mining process. The missing values may cause problems during both training and testing phases. Missing data forces classifiers to produce inaccurate results. Hence, suitable strategies should be adopted to handle the missing data.

4. Modelling:

This step plays the role in the application of data mining algorithm for the data to obtain a model or pattern.

5. Evaluate:

This step involves the evaluation of the data mining results using statistical analysis and visu-alization methods. The performance of the classi-fier is determined by evaluating the accuracy of the classifier.

6. Deployment:

This step involves the deployment of results of the data mining algorithm to improve the existing process or for a new situation.

Machine Learning Applications

* **Sentiment analysis:**

This is an application of Natural Language Processing (NLP) where the words of documents are converted to sentiments like happy, Sad and angry which are captured by emotions effectively.

For movie reviews or product reviews, five stars or one star are automatically attached using sentiment analysis programs.

* **Recommendation systems:**

These are systems that make personalized purchases possible.

Ex: Amazon recommends users to find related books or books bought by people who have the same taste like you. and Netflix suggests shows or related movies of your taste.

* **Voice assistants**

Products like Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants. They take speech commands and perform tasks. These chatbots are the result of machine learning technologies.

* Technologies like Google Maps and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

Table 1.4: Applications Survey Table

S.No.	Problem Domain	Applications
1.	Business	Predicting the bankruptcy of a business firm
2.	Banking	Prediction of bank loan defaulters and detecting credit card frauds
3.	Image Processing	Image search engines, object identification, image classification, and generating synthetic images
4.	Audio/Voice	Chatbots like Alexa, Microsoft Cortana. Developing chatbots for customer support, speech to text, and text to voice
5.	Telecommunication	Trend analysis and identification of bogus calls, fraudulent calls and its callers, churn analysis
6.	Marketing	Retail sales analysis, market basket analysis, product performance analysis, market segmentation analysis, and study of travel patterns of customers for marketing tours
7.	Games	Game programs for Chess, GO, and Atari video games
8.	Natural Language Translation	Google Translate, Text summarization, and sentiment analysis
9.	Web Analysis and Services	Identification of access patterns, detection of e-mail spams, viruses, personalized web services, search engines like Google, detection of promotion of user websites, and finding loyalty of users after web page layout modification
10.	Medicine	Prediction of diseases, given disease symptoms as cancer or diabetes. Prediction of effectiveness of the treatment using patient history and Chatbots to interact with patients like IBM Watson uses machine learning technologies.
11.	Multimedia and Security	Face recognition/identification, biometric projects like identification of a person from a large image or video database, and applications involving multimedia retrieval
12.	Scientific Domain	Discovery of new galaxies, identification of groups of houses based on house type/geographical location, identification of earthquake epicenters, and identification of similar land use

Understanding Data

Machine learning algorithms involves large data-sets. Hence, it is necessary to understand the data and datasets before applying machine learning algorithms.

What is Data?

All facts are data. In computer systems, bits encode facts present in numbers, text, images, audio and video.

Today, business organizations are accumulating vast and growing amounts of data of the order of gigabytes, tera bytes, exabytes.

1 byte = 8 bits

A bit is either 0 or 1.

1 KB (kilo byte) = 1024 bytes

1 MB (mega byte) = 1024 KB

1 GB (giga byte) ≈ 1,000,000 KB

1000 GB ≈ 1 TB (tera byte)

1000000 ≈ 1 EB (Exa byte)

1 Bit = Binary digit 0 or 1
8 Bits = 1 Byte
1024 Byte = 1 KB (Kilo Byte)
1024 KB = 1 MB (Mega Byte)
1024 MB = 1 GB (Giga Byte)
1024 GB = 1 TB (Tera Byte)
1024 TB = 1 PB (Peta Byte)
1024 PB = 1 EB (Exa Byte)
1024 EB = 1 ZB (Zetta Byte)
1024 ZB = 1 YB (Yotta Byte)
1024 YB = 1 BB (Bronto Byte)
1024 BB = 1 Geop Byte

* Data is available in different data sources like flat files, databases, or data warehouses.
* It can either be an operational data or a non-operational data.
* Operational data is the one that is encounted in normal business procedures & processes.
  Ex: daily sales data.

* Non-operational data is the kind of data that is used for decision making.
* Data by itself is meaningless. It has to be processed to generate any information.
* Processed data is called information that includes patterns, associations or relationships among data.

Ex: Sales data can be analyzed to extract information like which product was sold larger in the last quarter of the year.

Element of Big Data

Data whose volume is less and can be stored and processed by a small-scale computer is called 'small data'.

Big data, on the other hand, is a larger data whose volume is much larger than 'small data' and is characterized as follows.

1. Volume:

Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data. Big data is measured in terms of petabytes (PB) and exabytes (EB).

2. Velocity:

The fast arrival speed of data and its increase in data volume is noted as velocity. Velocity helps to understand the relative growth of big data & its accessibility by users, systems and applications.

3. Variety – The variety of Big Data includes

* Form – There are many forms of data. Data types range from text, graph, audio, video, to maps
* Function – These are data from various sources like human conversations, transaction records, and old archive data.
* Source of data – This is the third aspect of variety. There are many sources of data. Broadly, the data source can be classified as open / public data, social media data and multimodal data.

4. Veracity of data – Varacity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data. There are many sources of error such as technical errors, typographical errors, and human errors. So veracity is one of the most important aspects of data.

5. Validity – validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.

6. Value – Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

Thus, these 6 Vs are helpful to characterize the big data.

The data quality of the numerical attributes is determined by factors like precision, bias, and accuracy.

Precision is defined as the closeness of repeated measurements. Often, standard deviation is used to measure the precision.

Bias is a systematic result due to erroneous assumptions of the algorithms or procedures.

Accuracy is the degree of measurement of errors that refers to the closeness of measurements to the true value of the quantity.

Types of Data

In Big Data, there are three kinds of data. They are structured data, unstructured data, and semi-structured data.

* Structured Data

In structured data, data is stored in an organized manner such as a database where it is available in the form of a table. The data can also be retrieved in an organized manner using tools like SQL.

The structured data frequently encountered in machine learning are listed below.

=> Record Data:

A dataset is a collection of measurements taken from a process.

We have a collection of objects in a dataset and each object has a set of measurements. The measurements can be arranged in the form of a matrix.

Rows in the matrix represent an object and can be called as entities, cases, or records.

The columns of the dataset are called attributes, features, or fields.

The table is filled with observed data.

Label is the term that is used to describe the individual observations.

=> Data Matrix:

It is a variation of the record type because it consists of numeric attributes. The standard matrix operations can be applied on these data. The data is thought of as points or vectors in the multidimensional space where every attribute is a dimension describing the object

=> Graph data:

It involves the relationships among objects Ex: A web page can refer to another web page. This can be modeled as a graph. The nodes are web pages and the hyperlink is an edge that connects the nodes.

=> Ordered Data:

Ordered data objects involve attributes that have an implicit order among them.

The examples of ordered data are

1. Temporal data :-

It is the data whose attributes are associa-ted with time.

Ex: The customer purchasing patterns during festival time is Sequential data.

Time series data is a special type of sequence data where the data is a series of measurements over time.

2. Sequence data :-

It is like sequential data but does not have time stamps. This data involves the sequence of words or letters.

Ex: DNA data is a sequence of four characters -A T G C [Adenine(A), Thymine(T), Guanine(G), cytosine(C)]

3. Spatial data :-

It has attributes such as positions or areas

Ex: Maps are spatial data where the points are related by location.

* Unstructured Data:

Unstructured data includes video, image, and audio. It also includes textual documents, programs, and blog data.

It is estimated that 80% of the data are unstructured data.

* Semi-Structured Data

Semi-Structured data are partially unstructured. These include data like XML/JSON data, RSS feeds and hierarchical data.

Data Storage and Representation

Once the dataset is assembled, it must be stored in a structure that is suitable for data analysis. The goal of data storage management is to make data available for analysis. There are different approaches to organize and manage data in storage files and systems from flat file to data warehouse.

Flat files:

These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data. These flat files are the files where data is stored in plain ASCII or EBCDIC format.

Minor changes of data in flat files affect the results of the data mining algorithms. Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.

Spreadsheet formats

* **CSV files**: CSV stand for comma-separated values files where the values are separated by commas. These are used by spreadsheet and database applications.

The first row may have attributes and the rest of the rows represent the data.

* TSV files:

TSV files stands for Tab separated values files where values are separated by Tab.

Both CSV and TSV files are generic in nature and can be shared. There are many tools like Google sheets and Microsoft Excel to process these files.

Database System:

* It normally consists of database files and a database management system (DBMS).
* Database files contain original data & metadata.
* A relational database consists of sets of tables.
* The tables have rows and columns.
* The columns represent the attributes and rows represent tuples.
* A tuple corresponds to either an object or a relationship between objects.
* A user can access and manipulate the data using SQL.

Different types of databases

1. Transactional Databases:

A transactional database is a collection of transactional records. Each record is a transaction. A transaction may have a time stamp, identifier and a set of items, which may have links to other tables.

2. Time-Series database:

Time-Series database stores time related information like log files where data is associated with a time stamp. This data represents the sequence of data, which represent values or events obtained over a period (for example hourly, weekly or yearly) or repeated time span. Ex: Sales of a product.

3. Spatial databases:

Spatial databases contain spatial information in a raster or vector format.

Raster formats are either bitmaps or pixel maps. Ex: images can be stored as a raster data.

Vector format can be used to store maps as maps use basic geometric primitive like points, lines, polygons and so forth.

World Wide Web (WWW):

It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XML (eXtensible Markup Language):

It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.

Data Stream:

It is dynamic data, which flows in and out of the observing environment. Typical characte-ristics

of data stream are huge volume of data, dyna-mic, fixed order movement and real-time constraints.

Ex: Financial transaction data from ATMs and stock markets, social media feeds and web click stream data.

JSON (Javascript Object Notation)

It is another useful data interchange format that is often used for many machine learning algorithms.

Big Data Analytics and Types of Analytics

The primary aim of data analysis is to assist business organizations to take decisions.

Ex: A business organization may want to know which is the fastest selling product, in order for them to market activities.

Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.

Data analytics is a general term and data analysis is a part of it.

Data analytics is a general term and data analysis is a part of it.

Data analytics refers to the process of data collection, preprocessing and analysis. It deals with the complete cycle of data manage-ment. It helps in prediction.

There are four types of data analytics

1. Descriptive analytics
2. Diagnostic analytics
3. Predictive analytics
4. Prescriptive analytics

Descriptive Analytics :

It is about describing the main features of the data. After data collection is done, descriptive analytics deals with the collected data & quantifies it. Ex: calculating a student's GPA to summarize overall academic performance.

Diagnostic Analytics :

It deals with the question - why? This is also known as causal analysis, as it aims to find out the cause and effect of the events.

Ex: If a product is not selling, diagnostic ana -lytics aims to find out the reason.

Predictive Analysis :

It deals with the future. It deals with the ques -tion - what will happen in future given this data? This involves the application of algorithms to identify the patterns to predict the future. Ex: Amazon's Product Recommendations

Prescriptive Analytics :

It is about finding the best course of action for the business organizations. Prescriptive analytics goes beyond prediction and helps in decision making by giving a set of actions. Ex: Doctor prescribing a specific diet & medication based on a patient's test result to prevent future health issues.

Big Data Analysis Framework

For performing data analytics, many frameworks are proposed. All proposed analytics frameworks have some common factors.

Big data framework is a layered architecture. A 4-layer architecture has the following layers

1. Data connection layer
2. Data Management layer
3. Data Analytics layer
4. Presentation layer

Data connection Layer:

It has data ingestion mechanisms and data connectors. Data ingestion means taking raw data and importing it into appropriate data structures. It performs the tasks of ETL process. ETL means Extract, Transform and Load operations.

Data Management Layer:

It performs preprocessing of data. The purpose of this layer is to allow parallel execution of queries, and read, write and data management tasks.

Data Analytic Layer:

It has many functionalities such as statistical tests, machine learning algorithms to understand, and construction of machine learning models.

This layer implements many model validation mechanisms too.

Note:

Types of Processing

Cloud computing

Cloud computing is an emerging technology which is basically a business service model or simply called as pay-per-usage model.

The term 'Cloud' refers to the Internet that provides sharing of processing power, applications, storage and services.

It offers different kinds of services such as Iaas, Paas and Saas.

Saas (Software as a service) enables users to access software applications from the cloud.

Paas (Platform as a service) provides users the platform to develop and run their applications

Iaas (Infrastructure as a service) enables users to access the infrastructure required to run their applications, storage, operating systems, etc.

The cloud services can be deployed in four most commonly used deployment models

* Public cloud
* Private cloud
* Community Cloud
* Hybrid Cloud.

Public cloud: The public cloud is accessible to the public and is owned by a vendor, who offers the services of the cloud to the users.

Private cloud: Private cloud is a privately owned cloud where the user or an organization owns the cloud and only the user or employees of that organization have access to the cloud, thereby making data & transactions secure.

Community Cloud: In community cloud, the infrastructure is owned jointly by different organizations.

Hybrid cloud: The hybrid cloud is the combination of two or more cloud types.

Characteristics of cloud computing:

* Shared Infrastructure – Sharing of physical services, storage and networking capabilities
* Dynamic Provisioning – Resources assigned dynamically based on demands
* Dynamic Scaling – Expansion and contraction of service capability
* Network Access – Needs to be accessed across the internet.

* Utility based Metering - Uses metering to provide reporting and billing information
* Multitenancy - Serves multiple customers.
* Reliability - Customer reliable service.

Grid computing:

Grid computing is a parallel and distributed computing framework consisting of a network of computers offering a super computing service as a single virtual supercomputer.

It is required to perform specialized tasks that require a high computing power and a single computer cannot provide enough computing resources.

The grid computing model forms a grid by connecting tens of thousands of nodes as a cluster that runs on an operating system.

Grid is constructed by middleware software that evenly distributes the task to several nodes connected in the grid.

The individual nodes perform the task independently and in parallel which are then integrated to complete the large-scale task.

This model of computing is best suited for applications that are complex and can be computed in parallel.

H – Computing (High Performance Computing (HPC))

- It enables to perform complex tasks at high speed.
- It aggregates computing power in such a way that provides much higher performance to solve complex problems in science, engineering, research or business.
- It leverages parallel processing techniques for solving complex computational problems.
- HPC achieves this through concurrent use of computing resources.
- An HPC system combines the computing power of thousands of compute nodes that work in parallel to complete tasks faster.

Presentation Layer

It has mechanisms such as dashboards, and applications that display the results of analytical engines and machine learning algorithms.

The Big Data processing cycle involves

1. Data collection
2. Data preprocessing
3. Applications of Machine learning Algorithm
4. Interpretation of results & visualization of Machine learning algorithm.

Data Collection

The first task of gathering datasets are the collection of data. It is often estimated that most of the time is spent for collection of good quality data. A good quality data yields a better result.

'Good data' is one that has the following properties.

1. Timeliness – The data should be relevant and not stale or obsolete data.
2. Relevancy – The data should be relevant and ready for the machine learning or data mining algorithms. All the necessary information should be available and there should be no bias in the data.
3. Knowledge about the data – The data should be understandable & interpretable, and should be self-sufficient for the required application as desired by the domain knowledge engineer.

The data source can be classified as open/public data, social media data and multimodal data.

1. Open or public data source :-

It is a data source that does not have any stringent copyright rules or restrictions.

Ex: Government census data

& Digital libraries that have huge amount of text data as well as document images.

* Scientific domains with a huge collection of experimental data like genomic data and biological data.
* Healthcare systems that use extensive databases like patient databases, health insurance data, doctors information, and biomatrics information

2. Social Media –

It is the data that is generated by various social media platforms like Twitter, Facebook, Youtube, and Instagram. An enormous amount of data is generated by these platforms.

3. Multimodal data –

It includes data that involves many modes such as text, video, audio and mixed types.

* Image archives contain larger images databases along with numeric and text data.
* The World Wide Web (www) has huge amount of data that is distributed on the Internet.

Data Preprocessing

In real world, the available data is 'dirty'.

Dirty means.

* Incomplete data
* Outlier data
* Data with inconsistent values

* Inaccurate data
* Data with missing values
* Duplicate data.

Data preprocessing improves the quality of the data mining techniques.

The raw data must be preprocessed to give accurate results.

The process of detection and removal of errors in data is called data cleaning.

Data wrangling means making the data processable for machine learning algorithms.

The data errors include human errors such as typographical errors or incorrect measurement & structural errors like improper data formats. Data errors can also arise from omission and duplication of attributes.

Noise is a random component and involves distortion of a value or introduction of spurious objects.

Ex!

Patient ID	Name	Age	DOB	Fever	Salary
1	John	21		Low	-1500
2	Andre	36		High	Yes
3	David	5	10/10/1980	Low	" "
4	Raju	136		High	Yes

Illustration of 'Bad' Data.

It can be observed that data like Salary=" " is incomplete data.

The DoB of patients, John, Andre and Raju, is the missing data.

The age of David is recorded as '5' but his DoB indicates it is 10/10/1980. This is called inconsistent data.

Inconsistent data occurs due to problems in conversions, inconsistent formats, and difference in units. Salary for John is -1500. It cannot be less than '0'. It is an instance of noisy data.

Outliers are data that exhibit the characteristics that are different from other data and have very unusual values. The age of Raju cannot be 136. It might be typographical error.

Outliers may be legitimate data and sometimes are of interest to the data mining algorithms. These errors often come during data collection stage. These must be removed so that machine learning algorithms yield better results as the quality of results is determined by the quality of input data. This removal process is called data cleaning.

Missing Data Analysis

The primary data cleaning process is missing data analysis. Data cleaning routines attempt to fill up the missing values, smoothen the noise while identifying the outliers and correct the inconsistencies of the data. This enables data mining to avoid overfitting of the models.

The procedures that are given below can solve the problem of missing data.

1. Ignore the tuple :-
A tuple with missing data, especially the class label, is ignored. This method is not effective when the percentage of the missing values increases.

2. Fill in the values manually :
Here, the domain expert can analyze the data tables and carry out the analysis and fill in the values manually. But, this is time consuming and may not be feasible for larger sets.

3. A global constant can be used to fill in the missing attributes. The missing values may be 'Unknown' or be 'Infinity'. But, some data mining results may give spurious results by analysing these labels.

4. The attribute value may be filled by the average value. Ex: Average income can replace a missing value.

5. Use the attribute mean for all samples belonging to the same class. Here, the average value replaces the missing values of all tuples that fall in this group.

6. Use the most possible value to fill in the missing value. The most probable value can be obtained from other methods like classification and decision tree prediction.

Some of these methods introduce bias in the data. The filled value may not be correct and could be just an estimated value. Hence, the difference between the estimated & the original value is called an error or bias

Removal of Noisy or Outlier Data

Noise is a random error or variance in a measured value. It can be removed by using binning, which is a method where the given data values are sorted and distributed into equal frequency bins. The bins are also called as buckets. The binning method then uses the neigh-bor values to smooth the noisy data.

Binning Techniques:

* Smoothing by means – where the mean of the bin removes the values of the bins.
* Smoothing by bin medians – where the bin median replaces the bin values.
* Smoothing by bin boundaries – where the bin value is replaced by the closest bin boundary. The maximum and minimum values are called bin boundaries.

Example:

Consider the following set : S={12, 14, 19, 22, 24, 26, 28, 31, 32}. Apply various binning techniques and show the result.

By equal-frequency bin method, the data should be distributed across bins. Let us assume the bin of size 3, then the above data is distri-buted across the bins as shown below.

Bin 1 : 12, 14, 19

Bin 2 : 22, 24, 26

Bin 3 : 28, 31, 32

By smoothing bins method, the bins are replaced by the bin means.

Bin 1 : 15, 15, 15

Bin 2 : 24, 24, 24

Bin 3 : 30.3, 30.3, 30.3

Using smoothing by bin boundaries method, the bins values would be like.

Bin 1 : 12, 12, 19

Bin 2 : 22, 22, 26

Bin 3 : 28, 32, 32

As per the method, the minimum and maximum values of the bin are determined, and it serves as bin boundary and does not change. Rest of the values are transformed to the nearest value.

Data Integration and Data Transformations.

Data integration involves routines that merge data from multiple sources into a single data source. So, this may lead to redundant data.

The main goal of data integration is to detect and remove redundancies that arise from integration.

Data Transformation routines perform operations like normalization to improve the performance of the data mining algorithms. It is necessary to transform data so that it can be processed.

This can be considered as a preliminary stage of data conditioning.

Normalization is one such technique. In normalization, the attribute values are scaled to fit in a range (say 0-1) to improve the performance of the data mining algorithm.

Some of the normalization procedures used are

1. Min-Max
2. Z-Score

Min-Max Procedure

It is a normalization technique where each variable V is normalized by its difference with the minimum value divided by the range to a new range, say 0-1. The formula to implement this normalization is

$$\text{min-max} = \frac{V - \min}{\max - \min} \times (\text{new max} - \text{new min}) + \text{new min}$$

Here min-max is the range. min and max are the minimum and maximum of the given data, new max and new min are the minimum and maximum of the target range, say 0 and 1.

Example:

Consider the set: V = {88, 90, 92, 94}. Apply Min-Max procedure and map the marks to a new range 0–1.

Sol: The minimum of the list V is 88 and maximum is 94.

The new min and new max are 0 and 1.

$$\text{min-max} = \frac{V - \text{min}}{\text{max} - \text{min}} \times (\text{new max} - \text{new min}) + \text{new min}$$

For marks 88,

$$\text{min-max} = \frac{88-88}{94-88} \times (1-0) + 0 = 0$$

For marks 90,

$$\text{min-max} = \frac{90-88}{94-88} \times (1-0) + 0 = \frac{2}{6} = 0.33$$

For marks 92,

$$\text{min-max} = \frac{92-88}{94-88} \times (1-0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$\text{min-max} = \frac{94-88}{94-88} \times (1-0) + 0 = \frac{6}{6} = 1$$

The marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}.

Z-Score Normalization:

This procedure works by taking the difference between the field value and mean value, and by scaling this difference by standard deviation of the attribute.

$$V^* = V - \mu / \sigma$$

Here, σ is the standard deviation of the list V

μ is the mean of the list V.

Example:

Consider the mark list $V = \{10, 20, 30\}$, convert the marks to z-score.

Sol:

$$\text{(mean) } \mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\mu = \frac{10+20+30}{3} = 20$$

$$\text{(Standard deviation) } \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}}$$

$$\sigma = \sqrt{\frac{(10-20)^2 + (20-20)^2 + (30-20)^2}{3-1}}$$

$$\sigma = \sqrt{\frac{100 + 0 + 100}{2}} = \sqrt{\frac{200}{2}} = \sqrt{100}$$

$$\boxed{\sigma = 10}$$

$$\text{z-Score of } 10 = \frac{10-20}{10} = \frac{-10}{10} = -1$$

$$\text{z-Score of } 20 = \frac{20-20}{10} = \frac{0}{10} = 0$$

$$\text{z-Score of } 30 = \frac{30-20}{10} = \frac{10}{10} = 1$$

Hence, the z-Score of the marks 10, 20, 30 are $-1, 0, 1$, respectively.

What is the use of z-scores?

z-scores are used to detect outlier detection. If the data value z-score function is either less than −3 or greater than +3, then it is possibly an outlier.

The major disadvantage of z-score function is that it is extremely sensitive to outliers as it is dependent on mean.

Data Reduction

Data reduction reduces data size but produces the same results. There are different ways in which data reduction can be carried out such as data aggregation, feature selection, and dimensionality reduction.