UNIT 1

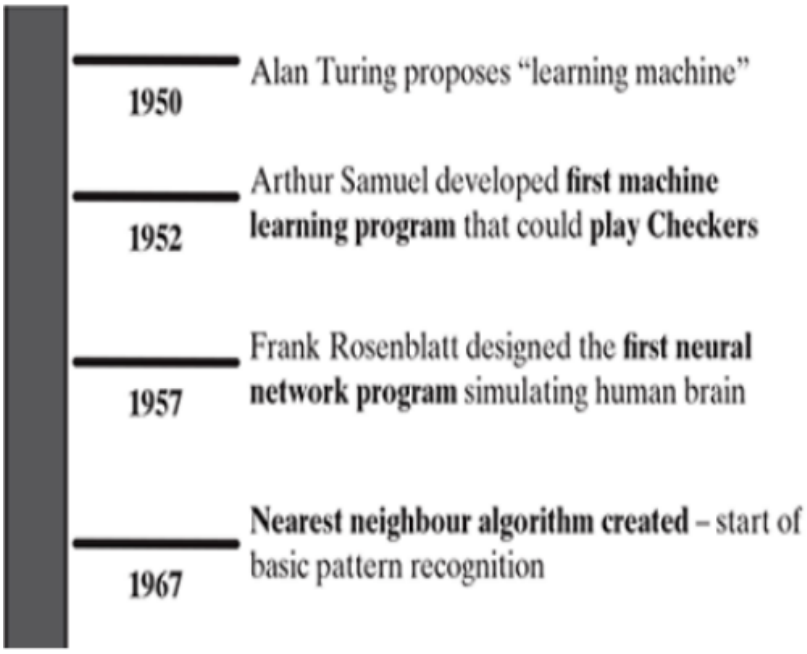
UNIT I Introduction to Machine Learning &Preparing to Model Lecture 9Hrs

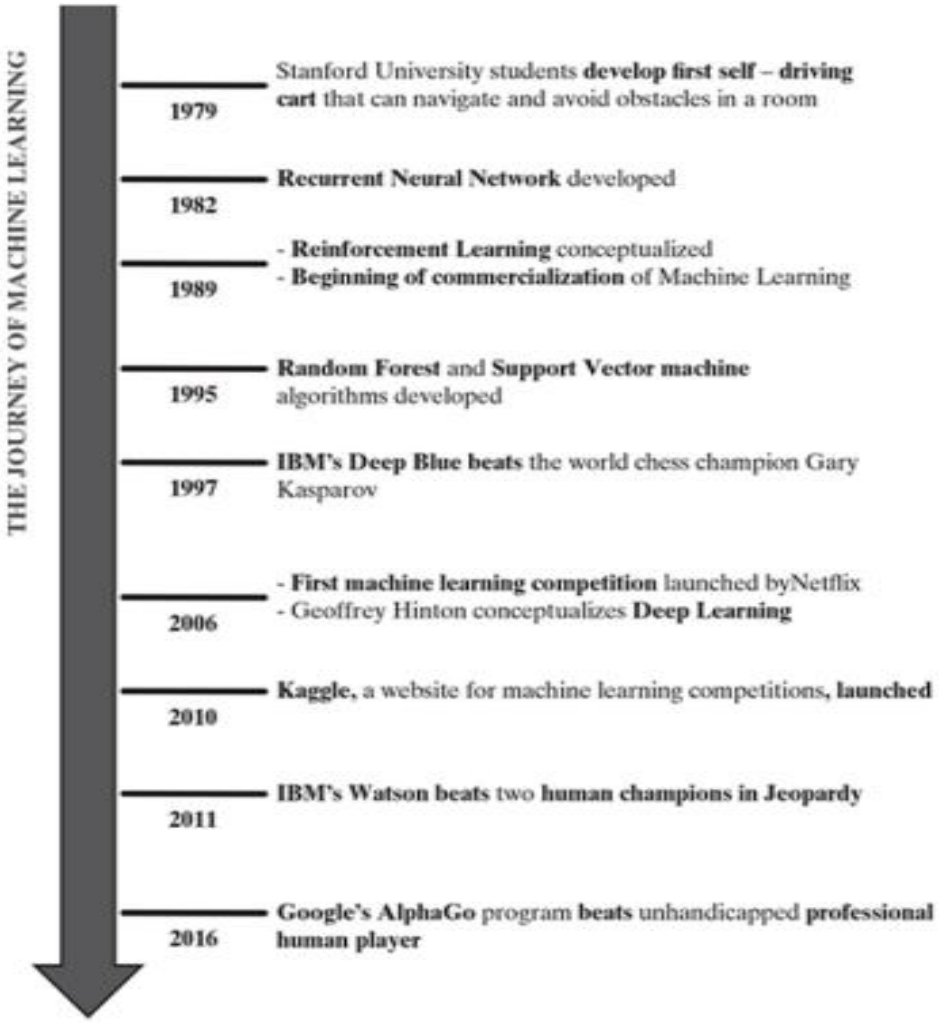
Introduction: What is Human Learning? Types of Human Learning, what is Machine Learning?Types of Machine Learning, Problems Not to Be Solved Using Machine Learning, Applications of Machine Learning, State-of-The-Art Languages/Tools in Machine Learning, Issues in Machine Learning

Preparing to Model: Introduction, Machine Learning Activities, Basic Types of Data in Machine Learning, Exploring Structure of Data, Data Quality and Remediation, Data Pre-Processing

1.1Evolution of Machine Learning:

- As of today, machine learning is a mature technology area finding its application in almost every sphere of life.
- It predicts the future market to help amateur traders compete with seasoned stock traders. It helps an oncologist find whether a tumour is malignant or benign.
- It helps in optimizing energy consumption thus helping the cause of Green Earth.
- Google has become one of the front-runners focusing a lot of its research on machine learning and artificial intelligence – Google self-driving car and Google Brain being two most ambitious projects of Google in its journey of innovation in the field of machine learning.
- In a nutshell, machine learning has become a way of life, no matter whichever sphere of life we closely look at.
- But where did it all start from?





1.2 WHAT IS HUMAN LEARNING?

- In cognitive science, learning is typically referred to as the process of gaining information through observation.
- In our daily life, we need to carry out multiple activities. It may be a task as simple as walking down the street or doing the homework. Or it may be some complex task like deciding the angle in which a rocket should be launched so that it can have a particular trajectory.
- To do a task in a proper way, we need to have prior information on one or more things related to the task. Also, as we keep learning more or in other words acquiring more information, the efficiency in doing the tasks keep improving.
- For example, with more knowledge, the ability to do homework with less number of mistakes increases.
- In the same way, information from past rocket launches helps in taking the right precautions and makes more successful rocket launch. Thus, with more learning, tasks can be performed more efficiently.

1.3 TYPES OF HUMAN LEARNING

- Thinking intuitively, human learning happens in one of the three ways

1. Either somebody who is an expert in the subject directly teaches us
2. we build our own notion indirectly based on what we have learnt from the expert in the past
3. we do it ourselves, may be after multiple attempts, some being unsuccessful.

- The first type of learning, we may call, falls under the category of learning directly under expert guidance, the second type falls under learning guided by knowledge gained from experts and the third type is learning by self or self learning.

1.3.1 Learning under expert guidance:

- As child we learn from parents, In schools the baby is able to learn all these things from his teacher who already has knowledge on these areas
- Then starts higher studies where the person learns about more complex, application-oriented skills. Engineering students get skilled in one of the disciplines like civil, computer science, electrical, mechanical, etc. medical students learn about anatomy, physiology, pharmacology, etc.
- Then the person starts working as a professional in some field. The professional mentors, by virtue of the knowledge that they have gained through years of hands-on experience, help all new comers in the field to learn on-job.
- In all phases of life of a human being, there is an element of guided learning. This learning is imparted by someone, purely because of the fact that he/she has already gathered the knowledge by virtue of his/her experience in that field. So guided learning is the process of gaining information from a person having sufficient knowledge due to the past experience.

1.3.2 Learning guided by knowledge gained from experts

- An essential part of learning also happens with the knowledge which has been told by teacher or mentor at some point of time in some other form/context.
- For example, a baby can group together all objects of same colour even if his parents have not specifically taught him to do so. He is able to do so because at some point of time or other his parents have told him which colour is blue, which is red, which is green, etc.
- A grown-up kid can select one odd word from a set of words because it is a verb and other words being all nouns. He could do this because of his ability to label the words as verbs or nouns, taught by his English teacher long back.
- In a professional role, a person is able to make out to which customers he should market a campaign from the knowledge about preference that was given by his boss long back.
- In all these situations, there is no direct learning. It is some past information shared on some different context, which is used as a learning to make decisions.

1.3.3 Learning by self:

- In many situations, humans are left to learn on their own.
- A classic example is a baby learning to walk through obstacles. He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle, he needs to cross over it. He faces the same challenge while learning to ride a cycle as a kid or drive a car as an adult. Not all things are taught by others.
- A lot of things need to be learnt only from mistakes made in the past. We tend to form a check list on things that we should do, and things that we should not do, based on our experiences

1.4 WHAT IS MACHINE LEARNING?

- The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as:

  Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.
- Tom M. Mitchell has defined machine learning as

  'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'

1.4.1 How do machines learn?

- The basic machine learning process can be divided into three parts.
  1. **Data Input:** Past data or information is utilized as a basis for future decision-making
  2. **Abstraction:** The input data is represented in a broader way through the underlying algorithm

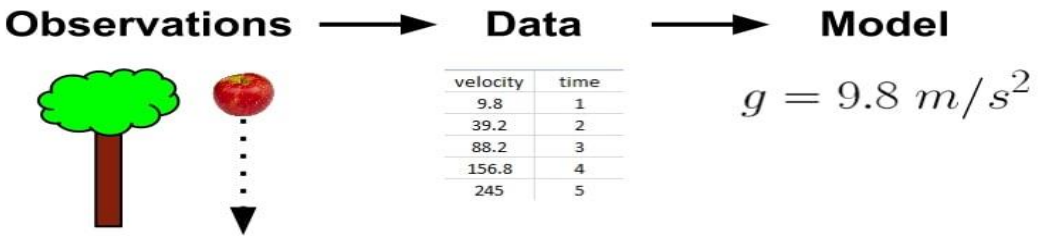
3. **Generalization:** The abstracted representation is generalized to form a framework for making decisions

FIG. 1.2 Process of machine learning

- Moving to the machine learning paradigm, the vast pool of knowledge is available from the data input.
- However, rather than using it in entirety, a concept map, this is nothing but knowledge abstraction as performed by the machine.
- In the end, the abstracted mapping from the input data can be applied to make critical conclusions. This is generalization in context of machine learning.

1.4.1.1 Abstraction:

- Representing raw input data in a structured format is the typical task for a learning algorithm. The work of assigning a meaning to data occurs during the **abstraction** process.
- During the process of knowledge representation, the computer summarizes raw inputs in a **model**
- The model may be in any one of the following forms
    - Computational blocks like if/else rules
    - Mathematical equations
    - Specific data structures like trees or graphs.
    - Logical groupings of similar observations
- The choice of model is typically not left up to the machine. Instead, the model is dictated by the learning task and the type of data being analysed.
- This process of fitting the model based on the input data is known as **training.** Also, the input data based on which the model is being finalized is known as **training data.**



1.4.1.2 Generalization:

- The first part of machine learning process is abstraction i.e. abstract the knowledge which comes as input data in the form of a model. However, this abstraction process, or more popularly training the model, is just one part of machine learning.
- The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions. This is achieved as a part of **generalization**.
- This part is quite difficult to achieve. This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics.
- But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems:

1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
2. The test data possess certain characteristics apparently unknown to the training data.

1.4.2 Well-posed learning problem:

- For defining a new problem, which can be solved using machine learning, a simple framework can be used. This framework also helps in deciding whether the problem is a right candidate to be solved using machine learning.
- The framework involves answering three questions:
  - **What is the problem?** Describe the problem informally and formally and list assumptions and similar problems.
  - **Why does the problem need to be solved?** List the motivation for solving the problem, the benefits that the solution will provide and how the solution will be used.
  - **How would I solve the problem?** Describe how the problem would be solved manually to flush domain knowledge.

1.5 TYPES OF MACHINE LEARNING(5 M):

- Machine learning can be classified into three broad categories:**(2 M)**
  1. **Supervised learning –** Also called predictive learning. A machine predicts the class of unknown objects based on prior class related information of similar objects.
  2. **Unsupervised learning –** Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects together.
  3. **Reinforcement learning –** A machine learns to act on its own to achieve the given goals.

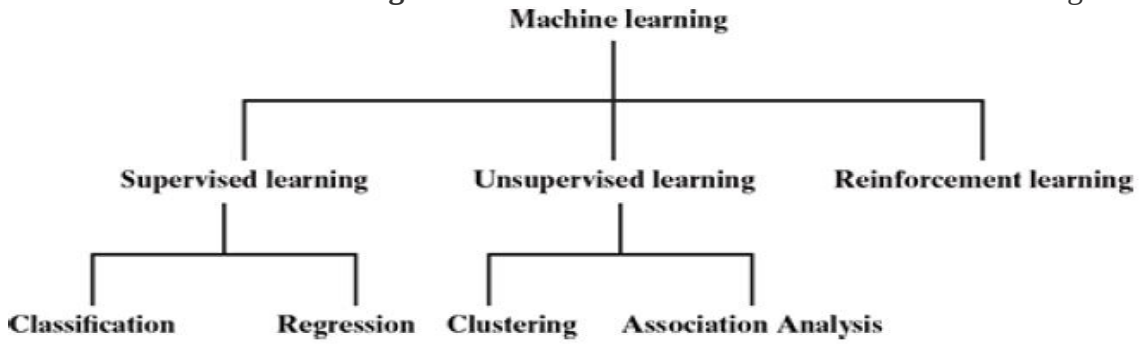


FIG. 1.3 Types of machine learning

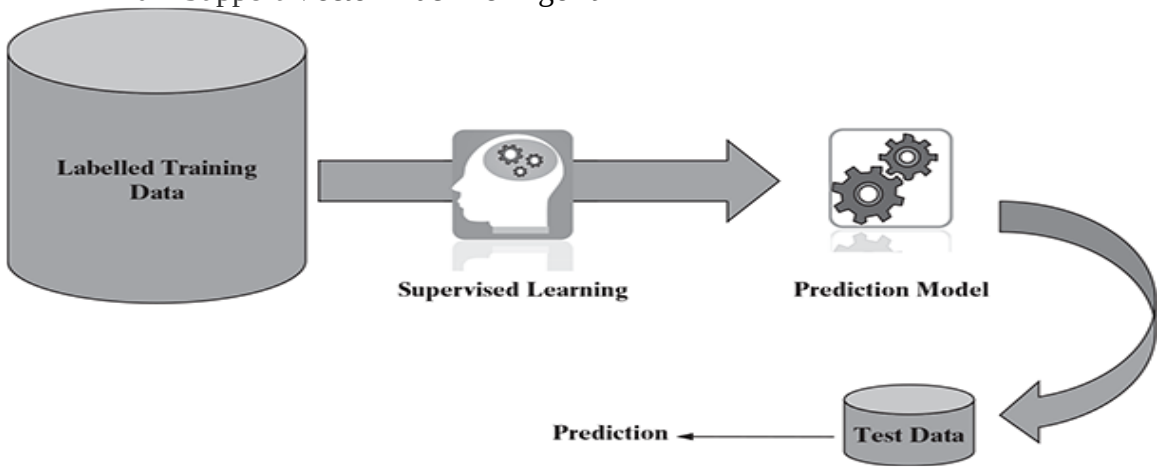
1.5.1 Supervised learning:

- In the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output.
- More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.
- The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.
- Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.
- Supervised machine learning can be classified into two types of problems, which are given below:

- **Classification**
- **Regression**

Classification:

- Classification is a type of supervised learning where a target feature, is predicted for test data based on the information given by training data. The target categorical feature is known as **class**.
- Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**.
- Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**
- Some popular classification algorithms are given below:
    - Random Forest Algorithm
    - Decision Tree Algorithm
    - Logistic Regression Algorithm
    - Support Vector Machine Algorithm



Regression:

- Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.
- Moreover, it is a type of supervised learning that learns from labelled data sets to predict continuous output for different data
- Some popular Regression algorithms are given below:
    - **Linear Regression Algorithm**
    - **Logistic Regression**
    - **Multivariate Regression Algorithm**

Advantages and Disadvantages of Supervised Learning

Advantages:

- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages:

- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

- **Image Segmentation** - Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.

- **Medical Diagnosis -** Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
- **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
- **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
- **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

1.5.2 Unsupervised learning:

- Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision
- In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or **patterns** within the data elements or records. Therefore, unsupervised learning is often termed as **descriptive model** and the process of unsupervised learning is referred as **pattern discovery** or **knowledge discovery**.
- One critical application of unsupervised learning is customer segmentation.

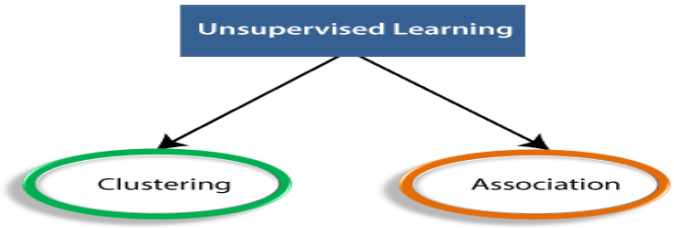
Why use Unsupervised Learning?

Below are some main reasons which describe the importance of Unsupervised Learning:

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Types of Unsupervised Learning Algorithm:

The unsupervised learning algorithm can be further categorized into two types of problems:



- **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
- **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

- **K-means clustering**

- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**
- **Singular value decomposition**

Advantages of Unsupervised Learning

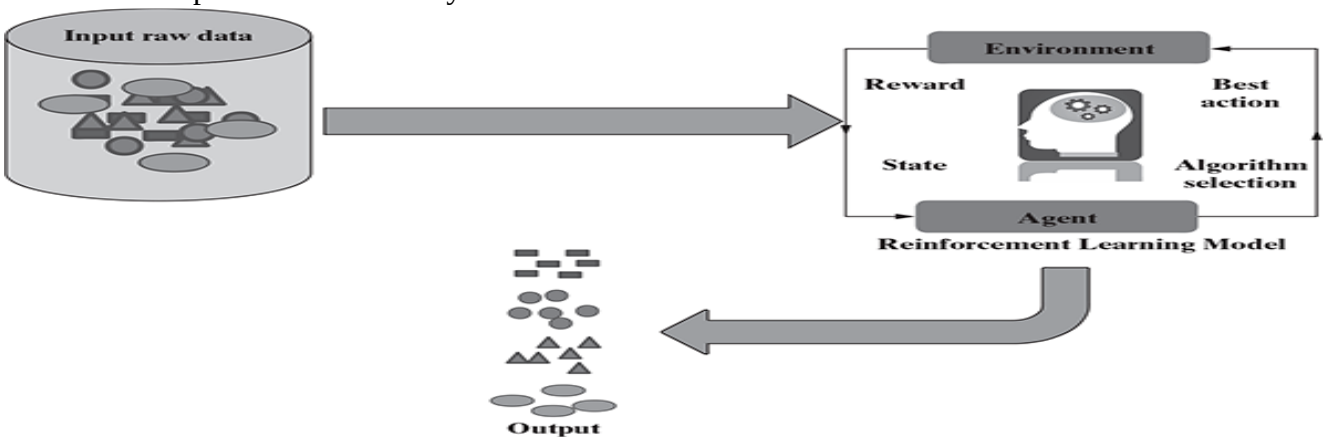
- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

1.5.3 Reinforcement learning:

- Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that
- Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.
- In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data
- It is a core part of Artificial intelligence, and all AI agent works on the concept of reinforcement learning. Here we do not need to pre-program the agent, as it learns from its own experience without any human intervention.



Reinforcement Learning Applications

1. **Robotics:**
    1. RL is used in **Robot navigation, Robo-soccer, walking, juggling**, etc.
2. **Control:**
    1. RL can be used for **adaptive control** such as Factory processes, admission control in telecommunication, and Helicopter pilot is an example of reinforcement learning.
3. **Game Playing:**

    1. RL can be used in **Game playing** such as tic-tac-toe, chess, etc.
4. **Chemistry:**
    1. RL can be used for optimizing the chemical reactions.
5. **Business:**
    1. RL is now used for business strategy planning.
6. **Manufacturing:**
    1. In various automobile manufacturing companies, the robots use deep reinforcement learning to pick goods and put them in some containers.
7. **Finance Sector:**
    1. The RL is currently used in the finance sector for evaluating trading strategies.

SUPERVISED	UNSUPERVISED	REINFORCEMENT
This type of learning is used when you know how to classify a given data, or in other words classes or labels are available.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification – it will get rewarded if the classification is correct, else get punished.
Labelled training data is needed. Model is built based on training data.	Any unknown and unlabelled data set is given to the model as input and records are grouped.	The model learns and updates itself through reward/ punishment.
The model performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values.	Difficult to measure whether the model did something useful or interesting. Homogeneity of records grouped together is the only measure.	Model is evaluated by means of the reward function after it had some time to learn.
There are two types of supervised learning problems – classification and regression.	There are two types of unsupervised learning problems – clustering and association.	No such types.
Simplest one to understand.	More difficult to understand and implement than supervised learning.	Most complex to understand and apply.
Standard algorithms include • Naïve Bayes • *k*-nearest neighbour (kNN) • Decision tree • Linear regression • Logistic regression • Support Vector Machine SVM), etc.	Standard algorithms are • *k*-means • Principal Component Analysis (PCA) • Self-organizing map (SOM) • Apriori algorithm • DBSCAN etc.	Standard algorithms are • Q-learning • Sarsa
Practical applications include • Handwriting recognition • Stock market prediction • Disease prediction • Fraud detection, etc.	Practical applications include • Market basket analysis • Recommender systems • Customer segmentation, etc.	Practical applications include • Self-driving cars • Intelligent robots • AlphaGo Zero (the latest version of DeepMind's AI system playing Go)

1.6 PROBLES NOT TO BE SOLVED USING MACHINE LEARNING

- Machine learning should not be applied to tasks in which humans are very effective or frequent human intervention is needed.
- For example, air traffic control is a very complex task needing intense human involvement.
- At the same time, for very simple tasks which can be implemented using traditional programming paradigms, there is no sense of using machine learning.
- For example, simple rule-driven or formula-based applications like price calculator engine, dispute tracking application, etc. do not need machine learning techniques.
- Machine learning should be used only when the business process has some lapses. If the task is already optimized, incorporating machine learning will not serve to justify the return on investment.
- For situations where training data is not sufficient, machine learning cannot be used effectively. This is because, with small training data sets, the impact of bad data is exponentially worse. For the quality of prediction or recommendation to be good, the training data should be sizeable.

1.7 APPLICATIONS OF MACHINE LEARNING:

- Wherever there is a substantial amount of past data, machine learning can be used to generate actionable insight from the data.
- Though machine learning is adopted in multiple forms in every business domain, we have covered below three major domains just to give some idea about what type of actions can be done using machine learning.

1.7.1 Banking and finance

- In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely prevalent. Since the volumes as well as velocity of the transactions are extremely high, high performance machine learning solutions are implemented by almost all leading banks across the globe.
- The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence.
- Customers of a bank are often offered lucrative proposals by other competitor banks. Also, sometimes customers get demotivated by the poor quality of services of the banks and shift to competitor banks.
- Machine learning helps in preventing or at least reducing the customer churn. Both descriptive and predictive learning can be applied for reducing customer churn. Using descriptive learning, the specific pockets of problem, i.e. a specific bank offering like car loan, may be spotted where maximum churn is happening. Using predictive learning, the set of vulnerable customers who may leave the bank very soon, can be identified. Proper action can be taken to make sure that the customers stay back.

1.7.2 Insurance

- Insurance industry is extremely data intensive. For that reason, machine learning is extensively used in the insurance industry.
- Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management.
- During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer.
- When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent.

1.7.3 Healthcare

- Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time. In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action.
- Suppose an elderly person goes for a morning walk in a park close to his house. Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the

wearable. The wearable data is sent to a remote server and a machine learning algorithm is constantly analyzing the streaming data. Alert can be sent to the person to immediately stop walking and take rest.

- Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

1.7.4: Image Recognition:

- Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**
- Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

1.7.5: Speech Recognition

- While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.
- Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

1.8 STATE-OF-THE-ART LANGUAGES/TOOLS IN MACHINE LEARNING

- The algorithms related to different machine learning tasks are known to all and can be implemented using any language/platform. However, there are certain languages and tools which have been developed with a focus for implementing machine learning. Few of them, which are most widely used, are covered below.

1.8.1 Python

- Python is one of the most popular, open source programming language widely adopted by machine learning community. It was designed by Guido van Rossum.
- Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms, mathematical tools (SciPy) and numerical plotting (matplotlib).
- Built on these libraries, there is a machine learning library named scikit-learn, which has various classification, regression, and clustering algorithms embedded in it.

1.8.2 R

- R is a language for statistical computing and data analysis. It is an open source language.
- R is considered as a variant of S, a GNU project which was developed at Bell Laboratories. Currently, it is supported by the R Foundation for statistical computing.
- R is a very simple programming language with a huge set of libraries available for different stages of machine learning. Some of the libraries standing out in terms of popularity are plyr/dplyr (for data transformation), caret (‘Classification and Regression Training’ for classification), RJava (to facilitate integration with Java), tm (for text mining), ggplot2 (for data visualization).

1.8.3 Matlab

- MATLAB (matrix laboratory) is a licenced commercial software with a robust support for a wide range of numerical computing. MATLAB is developed by MathWorks, a company founded in 1984.
- MATLAB also provides extensive support of statistical functions and has a huge number of machine learning algorithms in-built. It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

1.8.4 SAS

- SAS (earlier known as 'Statistical Analysis System') is another licenced commercial software which provides strong support for machine learning functionalities. Developed in C by SAS Institute, SAS had its first release in the year 1976.
- SAS is a software suite comprising different components. The basic data management functionalities are embedded in the Base SAS component whereas the other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

1.8.5 Other languages/tools

- There are a host of other languages and tools that also support machine learning functionalities. Owned by IBM, SPSS (originally named as Statistical Package for the Social Sciences) is a popular package supporting specialized data mining and statistical analysis. Originally popular for statistical analysis in social science (as the name reflects), SPSS is now popular in other fields as well.
- Released in 2012, Julia is an open source, liberal licence programming language for numerical analysis and computational science.

1.9 ISSUES IN MACHINE LEARNING

- Machine learning is a field which is relatively new and still evolving. Also, the level of research and kind of use of machine learning tools and technologies varies drastically from country to country.

1.9.1 privacy:

- The biggest fear and issue arising out of machine learning is related to privacy and the breach of it. The primary focus of learning is on analyzing data, both past and current, and coming up with insight from the data.
- This insight may be related to people and the facts revealed might be private enough to be kept confidential.
- For example, if there is a learning algorithm to do preference-based customer segmentation and the output of the analysis is used for sending targeted marketing campaigns, it will hurt the emotion of people and actually do more harm than good.
- So a very critical consideration before applying machine learning is that proper human judgement should be exercised before using any outcome from machine learning. Only then the decision taken will be beneficial and also not result in any adverse impact.

1.9.2: Inadequate Training Data

- The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data.
- Although data plays a vital role in the processing of machine learning algorithms, many data scientists claim that inadequate data, noisy data, and unclean data are extremely exhausting the machine learning algorithms.

1.9.3 Overfitting and Underfitting

Overfitting:

- Overfitting is one of the most common issues faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model.

Underfitting:

- Underfitting is just the opposite of overfitting. Whenever a machine learning model is trained with fewer amounts of data, and as a result, it provides incomplete and inaccurate data and destroys the accuracy of the machine learning model.

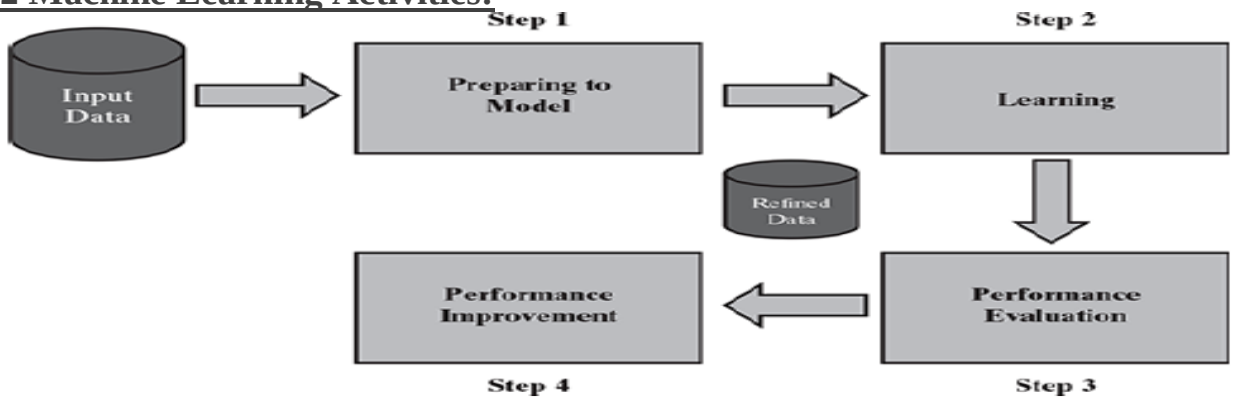
1.9.4 Data Bias

- Data Biasing is also found a big challenge in Machine Learning. These errors exist when certain elements of the dataset are heavily weighted or need more importance than others. Biased data leads to inaccurate results, skewed outcomes, and other analytical errors.

*****END*****

2. Preparing to Model

2.2 Machine Learning Activities:



Detailed process of machine learning

- The first step in machine learning activity starts with data. In case of supervised learning, it is the labelled training data set followed by test data which is not labelled. In case of unsupervised learning, there is no question of labelled data but the task is to find patterns in the input data.
- Multiple pre-processing activities may need to be done on the input data before we can go ahead with core machine learning activities.
- Following are the typical **preparation** activities done once the input data comes into the machine learning system:
    - Understand the type of data in the given input data set.
    - Explore the data to understand the nature and quality.
    - Explore the relationships amongst the data elements.
    - Find potential issues in data.Do the necessary remediation, e.g. impute missing data values, etc., if needed.
    - Apply pre-processing steps, as necessary.
- Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:
    - The input data is first divided into two parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
    - Consider different models or learning algorithms for selection.
    - Train the model based on the training data for supervised learning problem and apply to unknown data. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.
- After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

2.3 Basic Types of Data in Machine Learning

- A data set is a collection of related information or records.
- Let take a data set on student performance which has records providing performance, i.e. marks on the individual subjects. Each row of a data set is called a record. Each data set also has multiple attributes, each of which gives information on a specific characteristic

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

- let's try to understand the different types of data that we generally come across in machine learning problems. Data can broadly be divided into following two types:
  - **Qualitative data**
  - **Quantitative data**

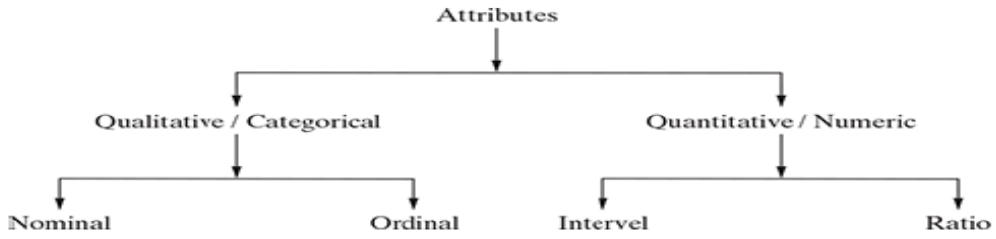


FIG. 2.4 Types of data

2.3.1 Qualitative data:

- **Qualitative data** provides information about the quality of an object or information which cannot be measured. For example, if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor', it falls under the category of qualitative data.
- Qualitative data is also called **categorical data**. Qualitative data can be further subdivided into two types as follows:
  - Nominal data
  - Ordinal data

Nominal data:

- **Nominal data** is one which has no numeric value, but a named value. It is used for assigning named values to attributes. Nominal values cannot be quantified.
- Examples of nominal data are
  1. Blood group: A, B, O, AB, etc.
  2. Nationality: Indian, American, British, etc.
  3. Gender: Male, Female, Other

Ordinal data:

- **Ordinal data**, in addition to possessing the properties of nominal data, can also be naturally ordered. This means ordinal data also assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value. Examples of ordinal data are
  1. Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
  2. Grades: A, B, C, etc.
  3. Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc.

2.3.2 Quantitative data:

- **Quantitative data** relates to information about the quantity of an object – hence it can be measured.
- For example, if we consider the attribute 'marks', it can be measured using a scale of measurement. Quantitative data is also termed as numeric data. There are two types of quantitative data:

1. Interval data
2. Ratio data

Interval data:

- **Interval data** is numeric data for which not only the order is known, but the exact difference between values is also known.
- An ideal example of interval data is Celsius temperature. The difference between each value remains the same in Celsius temperature.
- For example, the difference between 12°C and 18°C degrees is measurable and is 6°C as in the case of difference between 15.5°C and 21.5°C.
- Other examples include date, time, etc.
- For interval data, mathematical operations such as addition and subtraction are possible. For that reason, for interval data, the central tendency can be measured by mean, median, or mode. Standard deviation can also be calculated.

Ratio data:

- **Ratio data** represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, subtracted, multiplied, or divided. The central tendency can be measured by mean, median, or mode and methods of dispersion such as standard deviation. Examples of ratio data include height, weight, age, salary, etc.

2.4 EXPLORING STRUCTURE OF DATA

- Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

2.4.1 Exploring numerical data

- There are two most effective mathematical plots to explore numerical data – box plot and histogram. Starting with the most critical one, which is the box plot.

Understanding central tendency:

- To understand the nature of numeric variables, we can apply the measures of central tendency of data, i.e. mean and median
- In statistics, measures of central tendency help us understand the central point of a set of data.
- Mean, by definition, is a sum of all data values divided by the count of data elements. For example, mean of a set of observations – 21, 89, 34, 67, and 96 is calculated as below.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4$$

- Median, on contrary, is the value of the element appearing in the middle of an ordered list of data elements. If we consider the above 5 data elements, the ordered list would be – 21, 34, 67, 89, and 96. Since there are 5 data elements, the 3rd element in the ordered list is considered as the median. Hence, the median value of this set of data is 67.
- Mean being calculated from the cumulative sum of data values, is impacted if too many data elements are having values close to the maximum or minimum values. It is especially sensitive to **outliers,** i.e. the values which are unusually high or low, compared to the other values.

Understanding data spread:

- In some data sets, the data values are concentrated closely near the mean and in other data sets, the data values are more widely spread out from the mean. So, we will take a granular view of the data spread in the form of
    1. Dispersion of data
    2. Position of the different data values

1. **Measuring data dispersion:**
   - Consider the data values of two attributes
     1. Attribute 1 values : 44, 46, 48, 45, and 47
     2. Attribute 2 values : 34, 46, 59, 39, and 52
   - Both the set of values have a mean of 46.
   - However, the first set of values that is of attribute 1 is more concentrated or clustered around the mean/median value whereas the second set of values of attribute 2 is quite spread out or dispersed.
   - To measure the extent of dispersion of a data, or to find out how much the different values of a data are spread out, the variance of the data is measured. The variance of a data is measured using the formula given below:

$$\text{Variance}\ (x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2,$$

     - where x is the variable or attribute whose variance is to be measured and n is the number of observations
   - Standard deviation of a data is measured as follows:

$$\text{Standard deviation}\ (x) = \sqrt{\text{Variance}\ (x)}$$

   - Larger value of variance or standard deviation indicates more dispersion in the data and vice versa.
   - In the above example, let's calculate the variance of attribute 1 and that of attribute 2.
     For attribute 1

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

     For attribute 2,

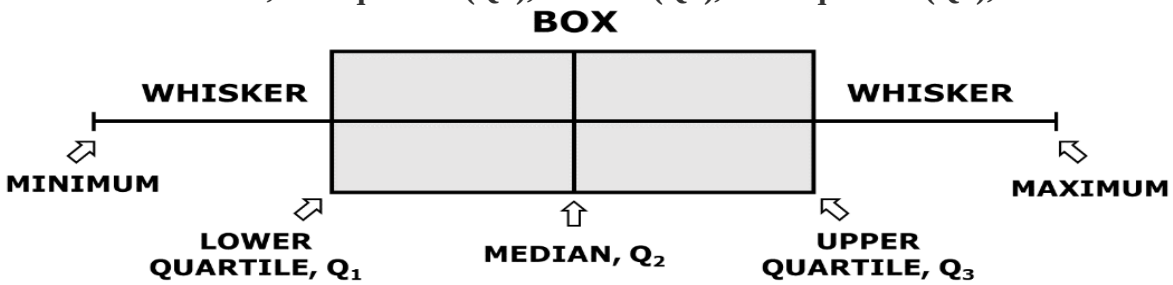
$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

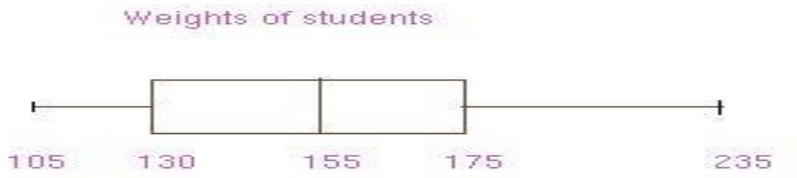
$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

   - So it is quite clear from the measure that attribute 1 values are quite concentrated around the mean while attribute 2 values are extremely spread out.
2. **Measuring data value position:**

- When the data values of an attribute are arranged in an increasing order, we have seen earlier that median gives the central data value, which divides the entire data set into two halves.
- Similarly, if the first half of the data is divided into two halves so that each half consists of one quarter of the data set, then that median of the first half is known as first quartile or Q_1 .
- In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q_3 .
- The overall median is also known as second quartile or Q_2. So, any data set has five values - **minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum**.



The five number summary is: 105,130, 155, 175, 235



2.4.2 Plotting and exploring numerical data

- There are two most effective mathematical plots to explore numerical data.
  - box plot
  - histogram

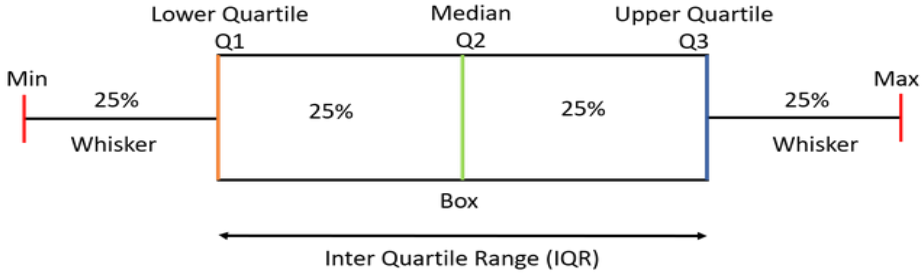
2.4.2.1 Box Plots:

Box Plot: It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

Components of a box plot

A box plot gives a five-number summary of a set of data which is-

- **Minimum** – It is the minimum value in the dataset excluding the outliers
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.



- The area inside the box (50% of the data) is known as the **Inter Quartile Range.** The **IQR** is calculated as –

IQR = Q3-Q1

- **Outliers** are the data points **below and above** the **lower and upper limit**. The lower and upper limit is calculated as –

  Lower Limit = Q1 - 1.5*IQR

  Upper Limit = Q3 + 1.5*IQR
- The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

How to create a box plot

- Let us take a sample data to understand how to create a box plot.
- Here are the runs scored by a cricket team in a league of 12 matches – 100,120,110,150,110,140,130,170,120,220,140,110.
- To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.
- Ascending Order -

  100,110,110,110,120,120,130,140,140,150,170,220
- Median (Q2) = (120+130)/2 = 125 ; Since there were even values
- To find the First Quartile we take the first six values and find their median.

  Q1 = (110+110)/2 = 110
- For the Third Quartile, we take the next six and find their median.

  Q3 = (140+150)/2 = 145

Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them.

- Now, we need to calculate the Inter Quartile Range.

  IQR = Q3-Q1 = 145-110 = 35
- We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

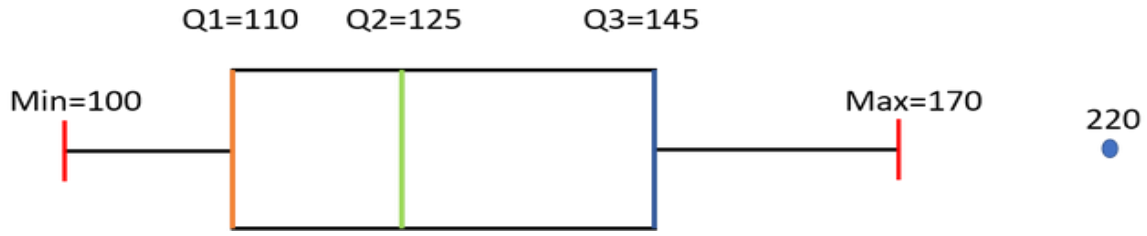
  Lower Limit = Q1-1.5*IQR = 110-1.5*35 = 57.5

  Upper Limit = Q3+1.5*IQR = 145+1.5*35 = 197.5
- So the minimum and maximum between the range [57.5,197.5] for our given data are

  Minimum = 100

  Maximum = 170
- The outliers which are outside this range are –

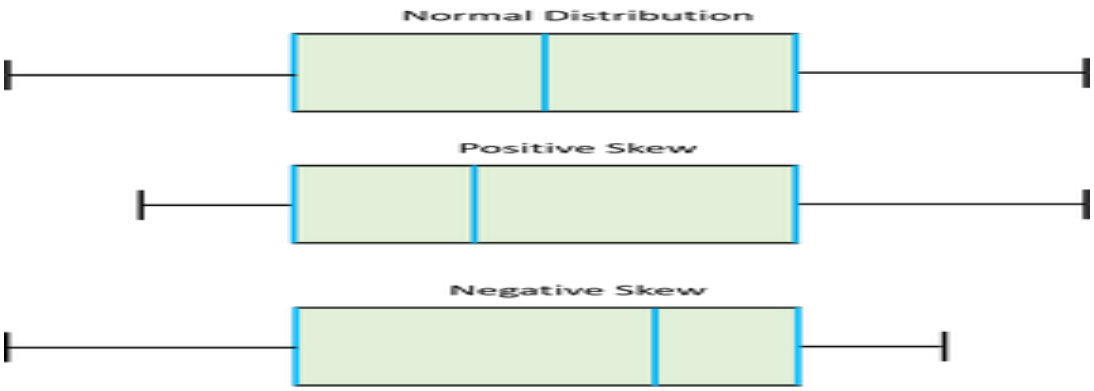
  Outliers = 220
- Now we have all the information, so we can draw the box plot which is as below-



- We can see from the diagram that the Median is not exactly at the centre of the box and one whisker is longer than the other. We also have one Outlier.

Uses of a Box Plot:

- Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).
- The Median gives you the average value of the data.
- Box Plots shows Skewness of the data-



- If the Median is at the center of the Box and the whiskers are almost the same on both the ends then the data is Normally Distributed.
- If the Median lies closer to the First Quartile and if the whisker at the lower end is shorter (as in the above example) then it has a Positive Skew (Right Skew)
- If the Median lies closer to the Third Quartile and if the whisker at the upper end is shorter then it has a Negative Skew (Left Skew).

- The dispersion or spread of data can be visualized by the minimum and maximum values which are found at the end of the whiskers.
- The Box plot gives us the idea of about the Outliers which are the points which are numerically distant from the rest of the data.

2.4.2.2 HISTOGRAM:

- Histogram is another plot which helps in effective visualization of numeric attributes. It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'.
- The important difference between histogram and box plot is
  - The focus of histogram is to plot ranges of data values (acting as 'bins'), the number of data elements in each range will depend on the data distribution. Based on that, the size of each bar corresponding to the different ranges will vary.
  - The focus of box plot is to divide the data elements in a data set into four equal portions, such that each portion contains an equal number of data elements.

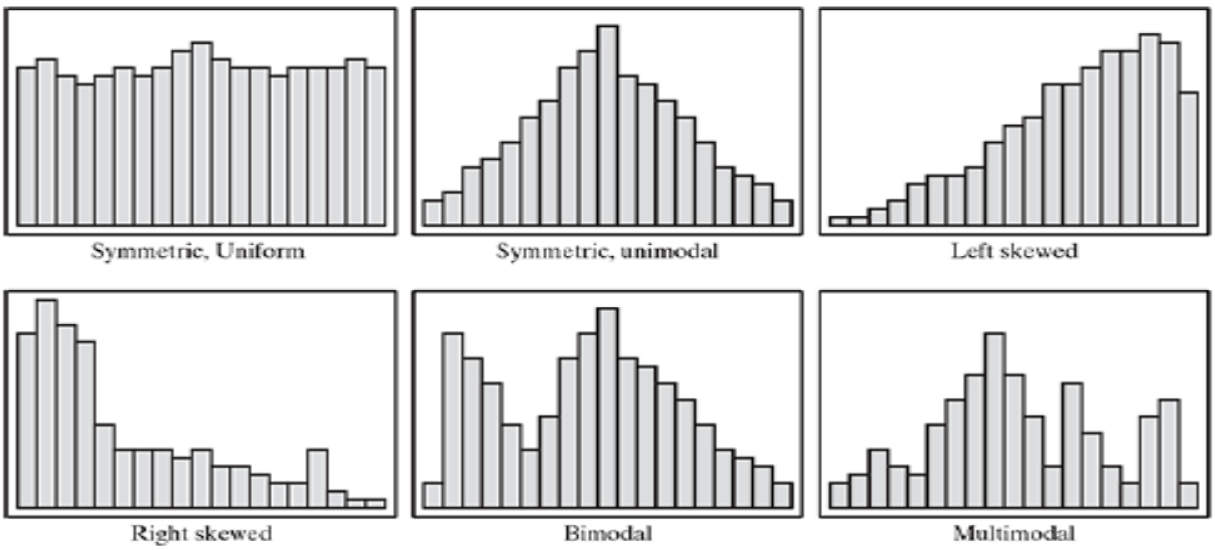
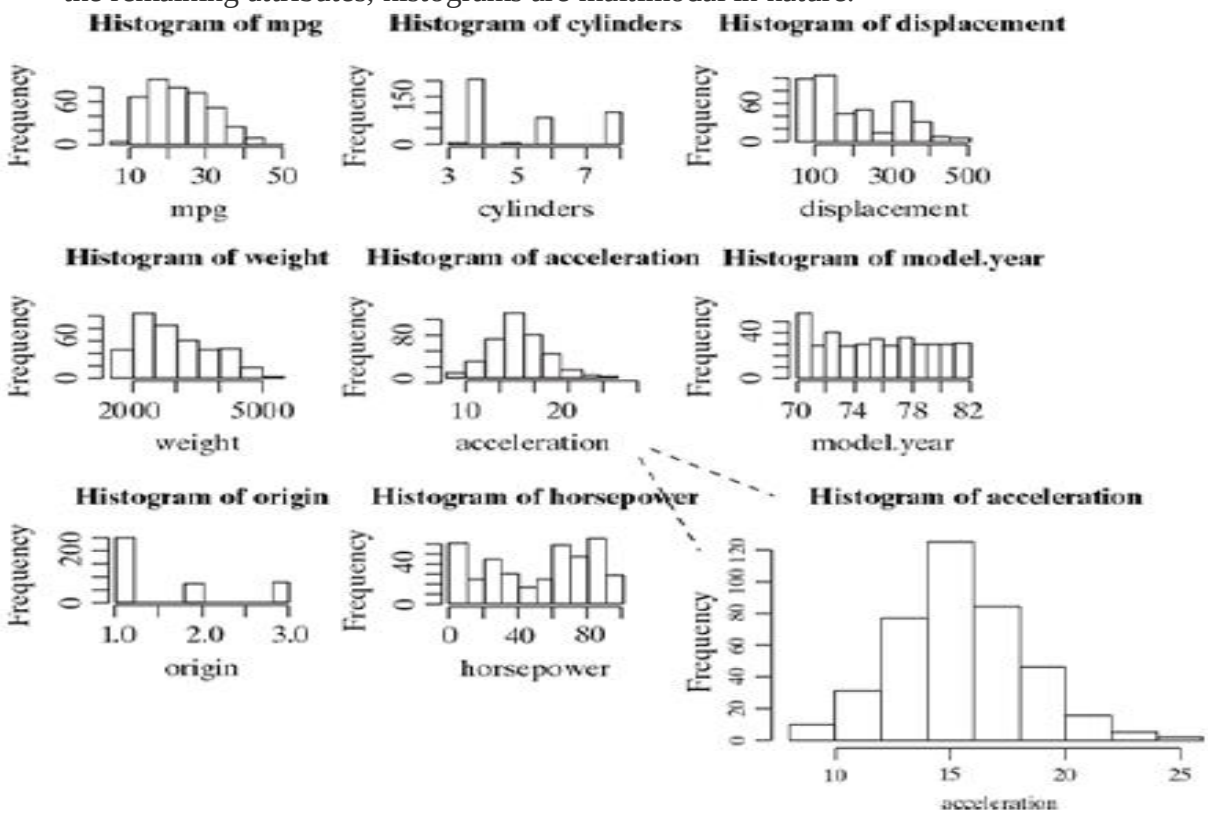


FIG. 2.11 General Histogram shapes

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500
14	8	454	220	4354	9	70	1	Chevrolet impala
14	8	440	215	4312	8.5	70	1	Plymouth fury iii
14	8	455	225	4425	10	70	1	Pontiac catalina
15	8	390	190	3850	8.5	70	1	Amc ambassador dpl
15	8	383	170	3563	10	70	1	Dodge challenger se
14	8	340	160	3609	8	70	1	Plymouth ' cuda 340
15	8	400	150	3761	9.5	70	1	Chevrolet monte carlo
14	8	455	225	3086	10	70	1	Buick estate wagon (sw)
24	4	113	95	2372	15	70	3	Toyota corona mark ii
22	6	198	95	2933	15.5	70	1	Plymouth duster
18	6	199	97	2774	15.5	70	1	Amc hornet

FIG. 2.5 Auto MPG data set

- Let's now examine the histograms for the different attributes of Auto MPG data set. The histograms for 'mpg' and 'weight' are right-skewed. The histogram for 'acceleration' is symmetric and unimodal, whereas the one for 'model.year' is symmetric and uniform. For the remaining attributes, histograms are multimodal in nature.



2.4.3 Exploring categorical data:

- We have seen there are multiple ways to explore numeric data. However, there are not many options for exploring categorical data.
- Lets take sample data set

S.No	Owner ID	Car Name
1	501	Maruti
2	502	Toyota
3	503	Maruti
4	504	Honda
5	505	MG
6	506	Honda
7	507	Maruti
8	508	Toyota
9	509	Maruti
10	510	KIA

- The first we need to find how many unique names are there for the attribute 'car name'. We can get this as follows:

  Maruti
  Toyota
  Honda
  MG
  KIA

- We may also look for a little more details and want to get a table consisting the categories of the attribute and count of the data elements falling into that category

Name	**Count**
Maruti	4
Toyota	2
Honda	2
MG	1
KIA	1

- In the same way, we may also be interested to know the proportion (or percentage) of count of data elements belonging to a category.
- Example for the attributes 'Car name', the proportion of data elements belonging to the category Maruti is 4 ÷ 10 = 0.4, i.e. 40%., Toyota is 20%, Honda is 20%, MG is 10%, KIA is 10%

2.4.4 Exploring relationship between variables:

- Analyzing and visualizing variables one at a time is not enough. To make various conclusions and analyses when performing exploratory data analysis, we need to understand how the variables in a dataset interact with respect to each other.
- There are numerous ways to analyze this relationship visually, one of the most common methods is the use of popular **scatterplots.**

2.4.4.1 Scatter plot:

- A scatterplot is one of the most common visual forms when it comes to comprehending the relationship between variables at a glance.
- In the simplest form, this is nothing but a plot of Variable A against Variable B: either one being plotted on the x-axis and the remaining one on the y-axis

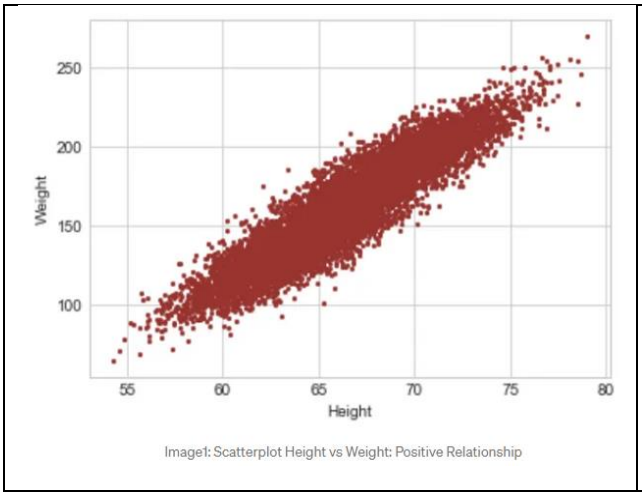


Image1: Scatterplot Height vs Weight: Positive Relationship

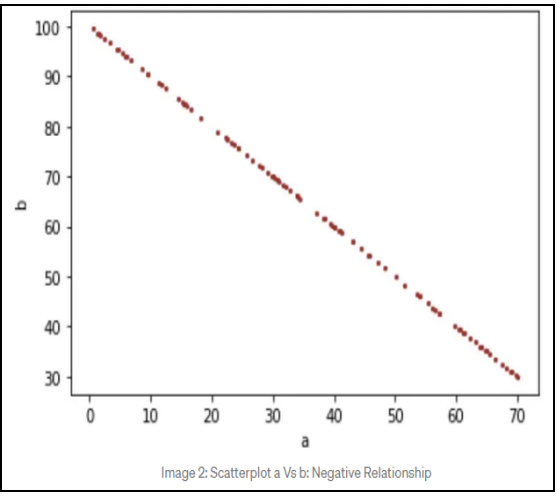


Image 2: Scatterplot a Vs b: Negative Relationship

2.4.4.2 Two-way cross-tabulations:

- Two-way cross-tabulations (also called cross-tab or contingency table) are used to understand the relationship of two categorical attributes in a concise way.
- It has a matrix format that presents a summarized view of the bivariate frequency distribution.
- A cross-tab, very much like a scatter plot, helps to understand how much the data values of one attribute changes with the change in data values of another attribute

2.5 DATA QUALITY AND REMEDIATION

2.5.1 Data quality

- Success of machine learning depends largely on the quality of data. A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning.
- We come across at least two types of problems:
    1. Certain data elements without a value or data are a missing value.
    2. Data elements having value surprisingly different from the other elements, which we term as outliers.
- There are multiple factors which lead to these data quality issues. Following are some of them:
    - **Incorrect sample set selection:** The data may not reflect normal or regular quality due to incorrect selection of sample set. For example, if we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future. In this case, the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time. It may also happen due to incorrect sample size. For example, a sample of small size may not be able to capture all aspects or information needed for right learning of the model.
- **Errors in data collection:** resulting in outliers and missing values In many cases, a person or group of persons are responsible for the collection of data to be used in a learning activity. In this manual process, there is the possibility of wrongly recording data in terms of value. This may result in data elements which have abnormally high or low value from other elements. Such records are termed as **outliers.** It may also happen that the data is not recorded at all. In case of a survey conducted to collect data, it is all the more possible as survey responders may choose not to respond to a certain question. So the data value for that data element in that responder's record is **missing.**

2.5.2 Data remediation

- The issues in data quality, as mentioned above, need to be remediated, if the right amount of efficiency has to be achieved in the learning activity. Let's see how to handle outliers and missing values.

2.5.2.1 Handling outliers:

- Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models.
- Once the outliers are identified and the decision has been taken to modify those values, you may consider one of the following approaches.
    - **Remove outliers:** If the number of records which are outliers is not many, a simple approach may be to remove them.
    - **Imputation:** One other way is to impute the value with mean or median or mode. The value of the most similar data element may also be used for imputation.
    - **Capping:** For values that lie outside the 1.5|×| IQR limits, we can cap them by replacing those observations below the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.
- If there is a significant number of outliers, they should be treated separately in the statistical model. In that case, the groups should be treated as two different groups, the model should be built for both groups and then the output can be combined.

2.5.2.2 Handling missing values:

- In a data set, one or more data elements may have missing values in multiple records. As discussed above, it can be caused by omission on part of the person who is collecting sample data
- There are multiple strategies to handle missing value of data elements. Some of those strategies have been discussed below.

1. Eliminate records having a missing value of data elements:

- In case the proportion of data elements having missing values is within a tolerable limit, a simple but effective approach is to remove the records having such data elements. This is possible if the quantum of data left after removing the data elements having missing values is sizeable.
- For Example in the case of Auto MPG data set, only in 6 out of 398 records, the value of attribute 'horsepower' is missing. If we get rid of those 6 records, we will still have 392 records, which is definitely a substantial number. So, we can very well eliminate the records and keep working with the remaining data set.
- However, this will not be possible if the proportion of records having data elements with missing value is really high as that will reduce the power of model

2. Imputing missing values:

- Imputation is a method to assign a value to the data elements having missing values. **Mean/mode/median** is most frequently assigned value.
- For **quantitative attributes**, all missing values are assigned with the mean, median, or mode of the remaining values under the same attribute.
- For **qualitative attributes**, all missing values are assigned by the mode of all remaining values of the same attribute.
- For example, in context of the attribute 'horsepower' of the Auto MPG data set, since the attribute is quantitative, we take a mean or median of the remaining data element values and assign that to all data elements having a missing value. So, we may assign the mean, which is 104.47 and assign it to all the six data elements.

3. Estimate missing values:

- If there are data points similar to the ones with missing attribute values, then the attribute values from those similar data points can be planted in place of the missing value.
- For finding similar data points or observations, distance function can be used.
- For example, let's assume that the weight of a Russian student having age 12 years and height 5 ft. is missing. Then the weight of any other Russian student having age close to 12 years and height close to 5 ft. can be assigned.

2.6 DATA PRE-PROCESSING

2.6.1 Dimensionality reduction

- High-dimensional data sets have high number of attributes or features that need a high amount of computational space and time. At the same time, not all features are useful – they degrade the performance of machine learning algorithms.
- Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. the number of features in the data set, is reduced.
- Dimensionality reduction helps in reducing irrelevance and redundancy in features. Also, it is easier to understand a model if the number of features involved in the learning activity is less.
- Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes.
- The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA). PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components.
- Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD).

2.6.2 Feature subset selection

- Feature subset selection or simply called feature selection, both for supervised as well as unsupervised learning, try to find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy.
- It may seem that a feature subset may lead to loss of useful information as certain features are going to be excluded from the final set of features used for learning.
- However, for elimination only features which are not relevant or redundant are selected.
- A feature is considered as irrelevant if it plays an insignificant role in classifying or grouping together a set of data instances. All irrelevant features are eliminated while selecting the final feature subset.
- A feature is potentially redundant when the information contributed by the feature is more or less same as one or more other features.
- Among a group of potentially redundant features, a small number of features can be selected as a part of the final feature subset without causing any negative impact to learn model accuracy.