HUMAN LEARNING

- Learning is typically referred to as the process of gaining information through observation.
- In daily life, multiple activities are needed to carry out. 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, prior information is needed on one or more things related to the task.
- Also, by learning more or by acquiring more information, the efficiency in doing the tasks keeps 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.

TYPES OF HUMAN LEARNING

- Human learning happens in one of the three ways –
- either somebody who is an expert in the subject directly teaches us, (learning directly under expert guidance)
- we build our own notion indirectly based on what we have learnt from the expert in the past, (learning guided by knowledge gained from experts)
- we do it ourselves, may be after multiple attempts, some being unsuccessful. (learning by self or selflearning)

Learning under expert guidance

- An infant may inculcate certain traits and characteristics, learning straight from its guardians/parents..
- The next phase of life is when the baby starts going to school. In school, he starts with basic familiarization of alphabets and digits. Then the baby learns how to form words from the alphabets and numbers from the digits.
- Slowly more complex learning happens in the form of sentences, paragraphs, complex mathematics, science, etc.
- 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.
- There are some experts, in general the teachers, in the respective field who have in-depth subject matter knowledge, who help the students in learning these skills.

- Then the person starts working as a professional in some field. Though he might have gone through enough theoretical learning in the respective field, he still needs to learn more about the hands-on application of the knowledge that he has acquired.
- 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.

Learning guided by knowledge gained from experts

- An essential part of learning also happens with the knowledge which has been imparted 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.

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.

MACHINE LEARNING

Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University. 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.'

In the context of the learning to play checkers,

- E represents the experience of playing the game,
- T represents the task of playing checkers and
- P is the performance measure indicated by the percentage of games won by the player.

The same mapping can be applied for any other machine learning problem, for example,image classification problem.

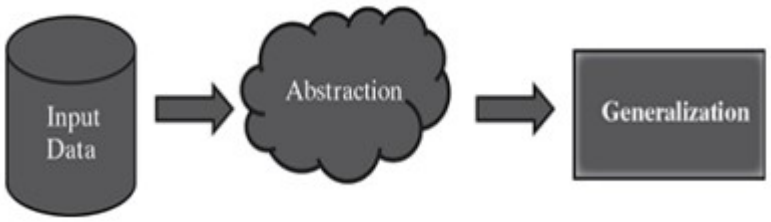
In context of image classification,

- E represents the past data with images having labels or assigned classes (for example whether the image is of a class cat or a class dog or a class elephant etc.),
- T is the task of assigning class to new, unlabelled images and
- P is the performance measure indicated by the percentage of images correctly classified.

The first step in any project is defining your problem. Even if the most powerful algorithm is used, the results will be meaningless if the wrong problem is solved.

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



Process of machine learning

Abstraction

- During the machine learning process, knowledge is fed in the form of input data.
- However, the data cannot be used in the original shape and form.
- Abstraction helps in deriving a conceptual map based on the input data.
- This map, or a model as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data.
- The choice of the model used to solve a specific learning problem is a human task.
- The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:

1. The type of problem to be solved:
2. Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
3. Nature of the input data:
4. Domain of the problem: If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

- Once the model is chosen, the next task is to fit the model based on the input data
- In a case where the model is represented by a mathematical equation, say 'y = c + c x' based on the input data, we have to find out the values of c and c . Otherwise, the equation (or the model) is of no use.
- So, fitting the model, in this case, means finding the values of the unknown coefficients or constants of the equation or the model
- 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.

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.
- 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:

  The trained model is aligned with the training data too much, hence may not portray the actual trend.

  The test data possess certain characteristics apparently unknown to the training data.
- Hence, a precise approach of decision-making will not work.
- An approximate or heuristic approach has to be adopted.
- This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality.
- But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

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:

1. What is the problem?

2. Why does the problem need to be solved?
3. How to solve the problem?

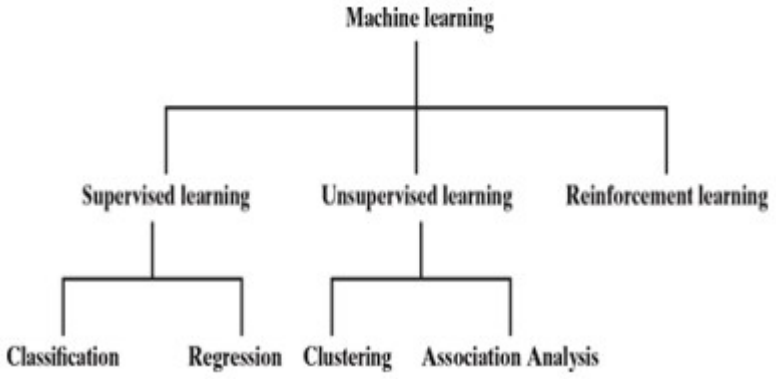
Step 1: What is the problem?
Describe the problem informally and formally and list assumptions and similar problems.
Informal description of the problem, e.g. I need a program that will prompt the next word as and when I type a word.
Tom Mitchell's machine learning formalism stated above to define the T, P, and E for the problem.
For example: Task (T): Prompt the next word when I type a word.
Experience (E): A corpus of commonly used English words and phrases.
Performance (P): The number of correct words prompted considered as a percentage (which in machine learning paradigm is known as learning accuracy).

Step 2: 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.

Step 3: How would I solve the problem?
Describe how the problem would be solved manually to flush domain knowledge.
Detail out step-by-step data collection, data preparation, and program design to solve the problem.
Collect all these details and update the previous sections of the problem definition, especially the assumptions

TYPES OF MACHINE LEARNING

Machine learning can be classified into three broad categories:
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.

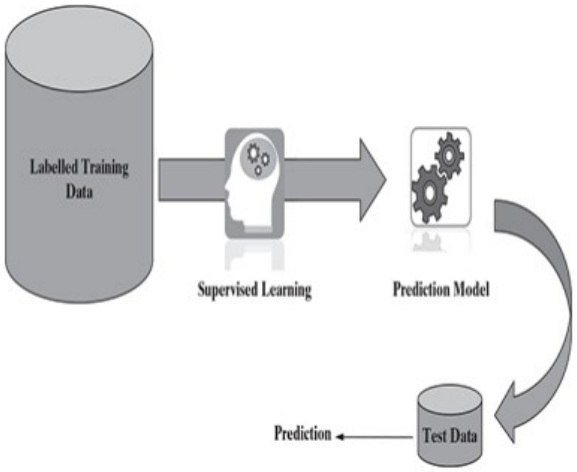


Supervised learning

- Supervised learning is a type of machine learning that uses labeled data to train machine learning models.
- In labeled data, the output is already known.
- The model just needs to map the inputs to the respective outputs.

For example.

- A machine is getting images of different objects as input and the task is to segregate the images by either shape or colour of the object.
- If it is by shape, the images which are of round-shaped objects need to be separated from images of triangular-shaped objects, etc.
- If the segregation needs to happen based on colour, images of blue objects need to be separated from images of green objects.
- But how can the machine can segregate For this a machine needs the basic information to be provided to it.
- This basic input, or the experience in the paradigm of machine learning, is given in the form of training data .
- Training data is the past information on a specific task. In context of the image segregation problem, training data will have past data on different aspects or features on a number of images, along with a tag on whether the image is round or triangular, or blue or green in colour.
- The tag is called ‘ label’ and we say that the training data is labelled in case of supervised learning.

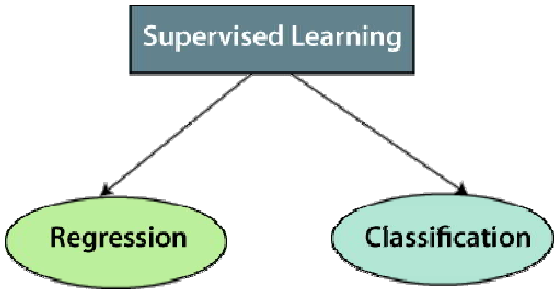


- Labelled training data containing past information comes as an input.
- Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

Some examples of supervised learning are

1. Predicting the results of a game
2. Predicting whether a tumour is malignant or benign
3. Predicting the price of domains like real estate, stocks, etc.
4. Classifying texts such as classifying a set of emails as spam or nonspam

Types of supervised Machine learning Algorithms:
Supervised learning can be further divided into two types of problems:



1. Regression
Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. Classification
Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.
Spam Filtering,

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

Advantages of Supervised learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages of supervised learning:

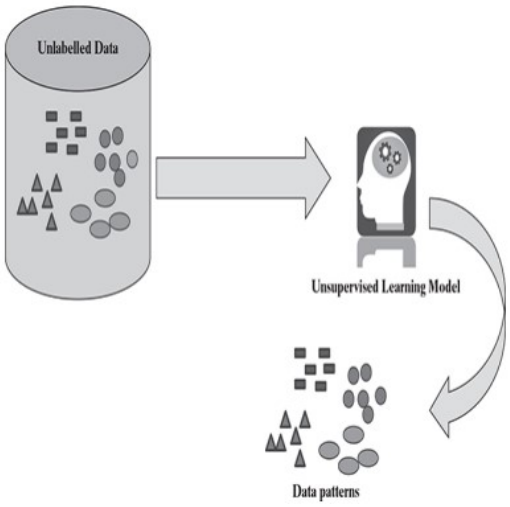
- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation times.
- In supervised learning, we need enough knowledge about the classes of object.

Unsupervised Learning

- unsupervised learning is a machine learning technique in which models are not supervised using training dataset.
- Instead, models itself find the hidden patterns and insights from the given data.
- It can be compared to learning which takes place in the human brain while learning new things.

It can be defined as:

- **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**.
- Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data.
- The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.



Example: Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the

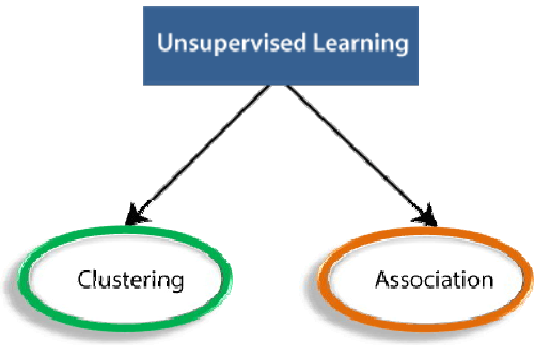
unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

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

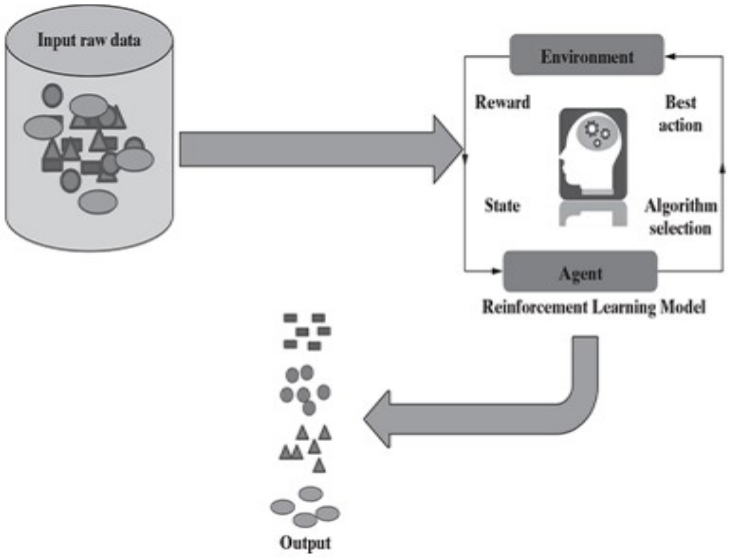
- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

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.

Reinforcement learning

- Machines often learn to do tasks autonomously.
- Let's try to understand in context of the example of the child learning to walk.
- The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment.
- It tries to improve its performance of doing the task.
- When a sub-task is accomplished successfully, a reward is given.
- When a sub-task is not executed correctly, obviously no reward is given.
- This continues till the machine is able to complete execution of the whole task.
- This process of learning is known as reinforcement learning



- One contemporary example of reinforcement learning is self-driving cars.
- The critical information which it needs to take care of are speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc.
- The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right, etc.

PROBLEMS 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.

APPLICATIONS OF MACHINE LEARNING

There are three major domains can be done using machine learning.

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.
- This helps in avoiding a lot of operational hassles in settling the disputes that customers will otherwise raise against those fraudulent transactions.
- Customers of a bank are often offered lucrative proposals by other competitor banks.
- Proposals like higher bank interest, lower processing charge of loans, zero balance savings accounts, no overdraft penalty, etc. are offered to customers, with the intent that the customer switches over to the competitor bank.
- 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 or a specific zone or a specific type of offering like car loan, may be spotted where maximum churn is happening.
- Quite obviously, these are troubled areas where further investigation needs to be done to find and fix the root cause.
- 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.

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.
- Other than the past information related to the specific customer, information related to similar customers, i.e. customer belonging to the same geographical location, age group, ethnic group, etc., are also considered to formulate the model.

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.
- In case of some extreme problem, doctors or healthcare providers in the vicinity of the person can be alerted.
- 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.
- It also has the history of the elderly person and persons of similar age group.
- The model predicts some fatality unless immediate action is taken.
- Alert can be sent to the person to immediately stop walking and take rest.
- Also, doctors and healthcare providers can be alerted to be on standby.
- Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

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

The algorithms related to different machine learning tasks are known to all and can be implemented using any language/platform.
It can be implemented using a Java platform or C / C++ language or in .NET.
However, there are certain languages and tools which have been developed with a focus for implementing machine learning.
Languages and tools, which are most widely used, are

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 and was first released in 1991.
- The reference implementation of Python, i.e. CPython, is managed by Python Software Foundation, which is a non-profit organization.
- Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms and 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.

R

- R is a language for statistical computing and data analysis. It is an open source language, extremely popular in the academic community – especially among statisticians and data miners.
- 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).
- Other than the libraries, certain packages like Shiny and R Markdown have been developed around R to develop interactive web applications, documents and dashboards on R without much effort.

Matlab

- MATLAB (matrix laboratory) is a licenced commercial software with a robust support for a wide range of numerical computing.
- MATLAB has a huge user base across industry and academia.
- MATLAB is developed by MathWorks, a company founded in 1984. Being proprietary software,
- MATLAB is developed much more professionally, tested rigorously, and has comprehensive documentation.
- 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.

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.

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 .
- 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.
- It has baked in all good things of MATLAB, Python, R, and other programming languages used for machine learning for which it is gaining steady attention from machine learning development community.
- Another big point in favour of Julia is its ability to implement high-performance machine learning algorithms.

ISSUES IN MACHINE LEARNING

- 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.
- Also, different people have a different preference when it comes to sharing of information.
- While some people may be open to sharing some level of information publicly, some other people may not want to share it even to all friends and keep it restricted just to family members.
- Classic examples are a birth date (not the day, but the date as a whole), photographs of a dinner date with family, educational background, etc.
- Some people share them with all in the social platforms like Facebook while others do not, or if they do, they may restrict it to friends only.
- When machine learning algorithms are implemented using those information, inadvertently people may get upset.
- 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.
- In certain countries, such events may result in legal actions to be taken by the people affected.
- Even if there is no breach of privacy, there may be situations where actions were taken based on machine learning may create an adverse reaction.
- Let's take the example of knowledge discovery exercise done before starting an election campaign.
- If a specific area reveals an ethnic majority or skewness of a certain demographic factor, and the campaign pitch carries a message keeping that in mind, it might actually upset the voters and cause an adverse result.
- 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.

MACHINE LEARNING ACTIVITIES

- 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.
- A thorough review and exploration of the data is needed to understand the type of the data, the quality of the data and relationship between the different data elements.
- Based on that, multiple pre-processing activities may need to be done on the input data before going with core machine learning activities.

The typical preparation activities done once the input data comes into the machine learning system:

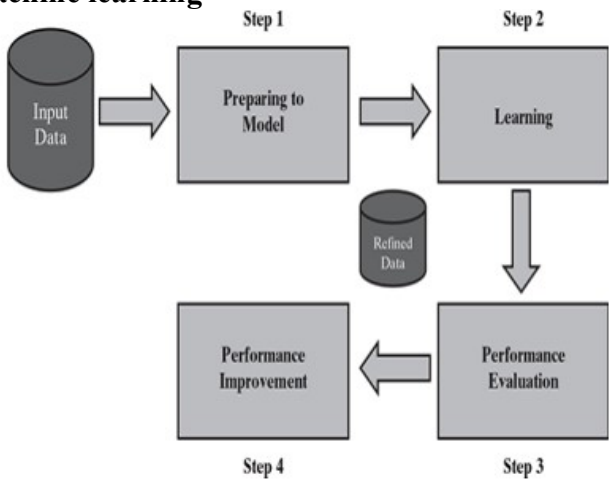
1. Understand the type of data in the given input data set.

2. Explore the data to understand the nature and quality.
3. Explore the relationships amongst the data elements, e.g. interfeature relationship.
4. Find potential issues in data.
5. Do the necessary remediation, e.g. impute missing data values, etc., if needed.
6. Apply pre-processing steps, as necessary.
7. Once the data is prepared for modelling, then the learning tasks start off.

As a part of it, do the following activities:

1. The input data is first divided into parts – the training data and the test data (called holdout).
2. This step is applicable for supervised learning only.
3. Consider different models or learning algorithms for selection.
4. Train the model based on the training data for supervised learning problem and applies to unknown data.
5. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.
6. After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated.
7. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

Four-step process of machine learning



Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

BASIC TYPES OF DATA IN MACHINE LEARNING

- A data set is a collection of related information or records.
- The information may be on some entity or some subject area.
- For example, in a data set on students in which each record consists of information about a specific student.
- 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.
- For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age, each of which understandably is a specific characteristic about the student entity.
- Attributes can also be termed as feature, variable, dimension or field.
- The data sets, Student, are having four features or dimensions; hence they are told to have four dimensional data space.
- A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features.
- Value of an attribute, may vary from record to record.

Data can broadly be divided into following two types:
1. Qualitative data
2. Quantitative 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.

Also, name or roll number of students are information that cannot be measured using some scale of measurement.

So they would fall under qualitative data.

Qualitative data is also called categorical data.

Qualitative data can be further subdivided into two types as follows:
1. Nominal data
2. Ordinal 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,

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.

Like nominal data, basic counting is possible for ordinal data.

Hence, the mode can be identified.

Since ordering is possible in case of ordinal data, median, and quartiles can be identified in addition.

Mean can still not be calculated.

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

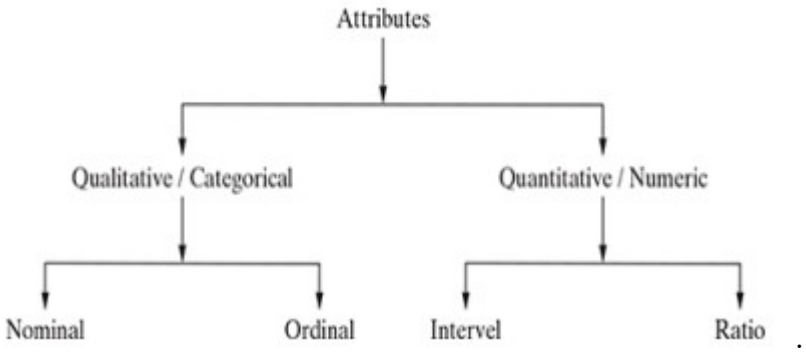
2. Ratio data

Interval data 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.
- However, interval data do not have something called a 'true zero' value.
- For example, there is nothing called '0 temperature' or 'no temperature'.
- Hence, only addition and subtraction applies for interval data.
- The ratio cannot be applied.
- This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C. However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

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.



- Attributes can also be categorized into types based on a number of values that can be assigned. The attributes can be either discrete or continuous based on this factor.
- Discrete attributes can assume a finite or countably infinite number of values.
- Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values whereas numeric attributes such as count, rank of students, etc. can have countably infinite values.
- A special type of discrete attribute which can assume two values only is called binary attribute.

- Examples of binary attribute include male/ female, positive/negative, yes/no, etc.
- Continuous attributes can assume any possible value which is a real number. Examples of continuous attribute include length, height, weight, price, etc.

EXPLORING STRUCTURE OF DATA

- The approach of exploring numeric data is different than the approach of exploring categorical data.
- We need to understand that in a data set, which of the attributes are numeric and which are categorical in nature
- In case of a standard data set, we may have the data dictionary available for reference.
- Data dictionary is a metadata repository, i.e. the repository of all information related to the structure of each data element contained in the data set.
- The data dictionary gives detailed information on each of the attributes – the description as well as the data type and other relevant details.
- In case the data dictionary is not available, we need to use standard library function of the machine learning tool that we are using and get the details.

Example dataset

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 acbassador

- The attributes such as ‘mpg’, ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’, ‘acceleration’, ‘model year’, and ‘origin’ are all numeric.
- Out of these attributes, ‘cylinders’, ‘model year’, and ‘origin’ are discrete in nature as the only finite number of values can be assumed by these attributes.
- The remaining of the numeric attributes, i.e. ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ can assume any real value.
- Hence, these attributes are continuous in nature.

- The only remaining attribute 'car name' is of type categorical, or more specifically nominal. This data set is regarding prediction of fuel consumption in miles per gallon, i.e. the numeric attribute 'mpg' is the target attribute.

With this understanding of the data set attributes, we can start exploring the numeric and categorical attributes separately

Exploring numerical data

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

1.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, 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 closer to the far end of the range, i.e. 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.
- Mean is likely to get shifted drastically even due to the presence of a small number of outliers.
- If we observe that for certain attributes the deviation between values of mean and median are quite high, we should investigate those attributes further and try to find out the root cause along with the need for remediation.

Example

	mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin
Median	23	4	148.5	?	2804	15.5	76	1
Mean	23.51	5.455	193.4	?	2970	15.57	76.01	1.573
Deviation	2.17	26.67%	23.22%		5.59%	0.45%	0.01%	36.43%
	Low	High	High		Low	Low	Low	High

- The comparison between mean and median for all the attributes has been shown.
- We can see that for the attributes such as 'mpg', 'weight', 'acceleration', and 'model.year' the deviation between mean and median is not significant which means the chance of these attributes having too many outlier values is less.
- However, the deviation is significant for the attributes 'cylinders', 'displacement' and 'origin'. So, we need to further drill down and look at some more statistics for these attributes.
- Also, there is some problem in the values of the attribute 'horsepower' because of which the mean and median calculation is not possible.

mpg	cylinders	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord di

Here, the 6 data elements, do not have value for the attribute 'horsepower'.
For that reason, the attribute 'horsepower' is not treated as a numeric.
That's why the operations applicable on numeric variables, like mean or median, are failing.
So we have to first remediate the missing values of the attribute 'horsepower' before being able to do any kind of exploration

2.Understanding data spread

- The central tendency of the different numeric attributes,gives a clear idea of which attributes have a large deviation between mean and median.
- To drill down more, we need to look at the entire range of values of the attributes, a granular view of the data spread in the form of

1. Dispersion of data
2. Position of the different data values

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 and median 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 or values of variable x.

- 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.

Measuring data value position

- When the data values of an attribute are arranged in an increasing order, 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 .
- 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 .
- The overall median is also known as second quartile or Q . So, any data set has five values minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

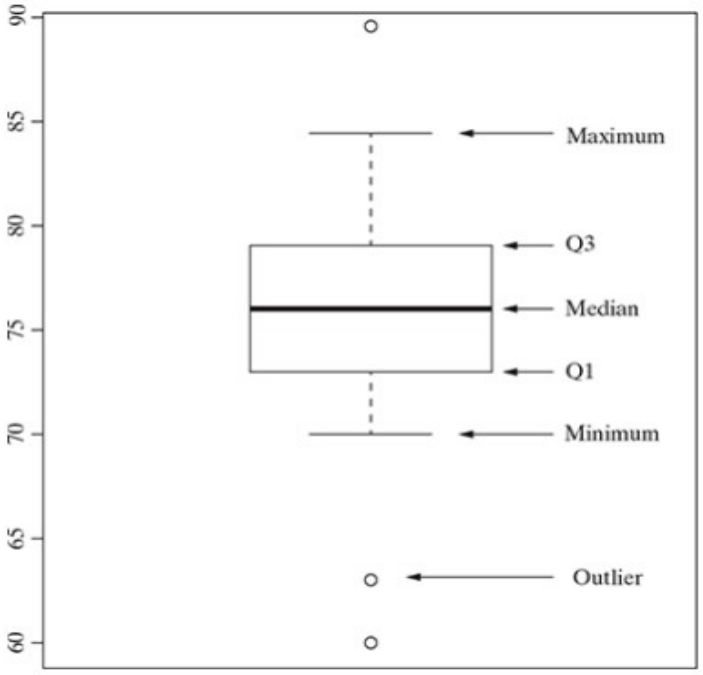
- In the example of the attribute ‘displacement’, we can see that the difference between minimum value and Q1 is 36.2 and the difference between Q1 and median is 44.3.
- The difference between median and Q3 is 113.5 and Q3 and the maximum value is 193.
- In other words, the larger values are more spread out than the smaller ones.
- This helps in understanding why the value of mean is much higher than that of the median for the attribute ‘displacement’.
- Similarly, in case of attribute ‘cylinders’, we can observe that the difference between minimum value and median is 1 whereas the difference between median and the maximum value is 4.
- For the attribute ‘origin’, the difference between minimum value and median is 0 whereas the difference between median and the maximum value is 2

- However, we still cannot ascertain whether there is any outlier present in the data. For that, we can better adopt some means to visualize the data. Box plot is an excellent visualization medium for numeric data

Quantiles refer to specific points in a data set which divide the data set into equal parts or equally sized quantities. There are specific variants of quantile, the one dividing data set into four parts being termed as quartile. Another such popular variant is percentile, which divides the data set into 100 parts

Plotting and exploring numerical data
Box plots

- A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data.
- Box plot (also called box and whisker plot) gives a standard visualization of the five number summary statistics of a data, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

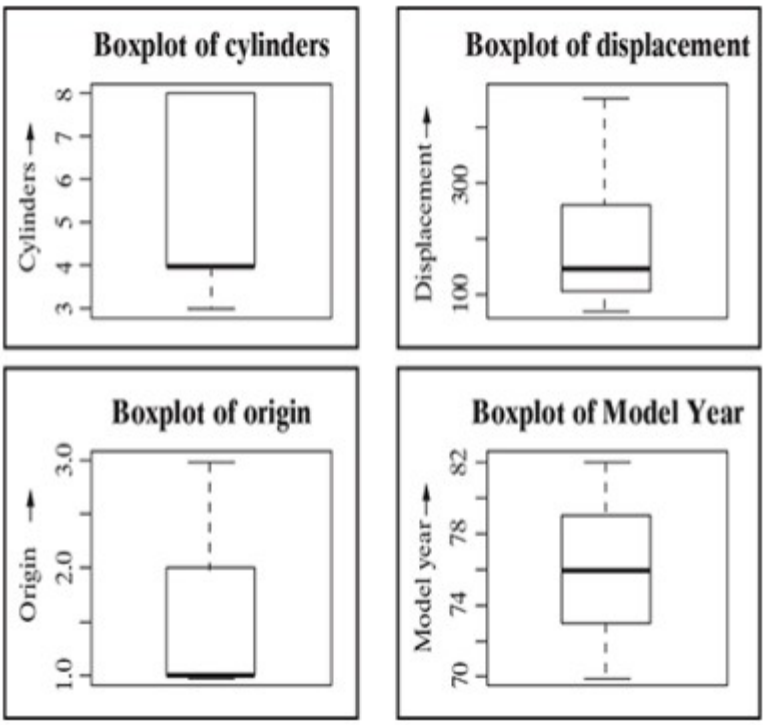


- The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving the inter-quartile range (IQR).
- Median is given by the line or band within the box.
- The lower whisker extends up to 1.5 times of the inter-quartile range (or IQR) from the bottom of the box, i.e. the first quartile or Q1.
- However, the actual length of the lower whisker depends on the lowest data value that falls within (Q1 − 1.5 times of IQR).

Example

- For a specific set of data, Q1 = 73, median = 76 and Q3 = 79.
- Hence, IQR will be 6 (i.e. Q3 – Q1)

- So, lower whisker can extend maximum till (Q1 – 1.5 × IQR) = 73 – 1.5 × 6 = 64.
- However, there are lower range data values such as 70, 63, and 60.
- So, the lower whisker will come at 70 as this is the lowest data value larger than 64.
- The upper whisker extends up to 1.5 as times of the inter-quartile range (or IQR) from the top of the box, i.e. the third quartile or Q3.
- Similar to lower whisker, the actual length of the upper whisker will also depend on the highest data value that falls within (Q3 + 1.5 times of IQR).
- Upper whisker can extend maximum till (Q3 + 1.5 × IQR) = 79 + 1.5 × 6 = 88.
- If there is higher range of data values like 82, 84, and 89.
- So, the upper whisker will come at 84 as this is the highest data value lower than 88.
- The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively.
- These are the outliers, which may deserve special consideration



The box plot for the three attributes - ‘cylinders’, ‘displacement’, and ‘origin

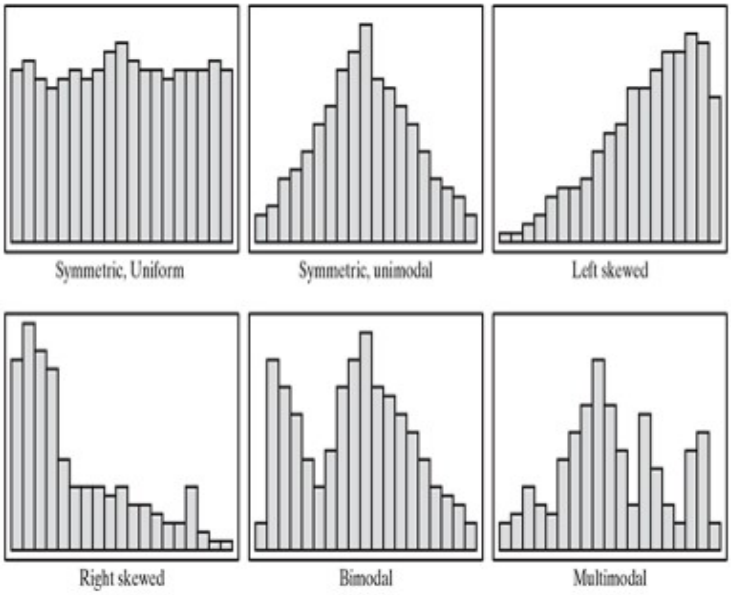
Table 2.2 *Frequency of "Cylinders" Attribute*

Cylinders	Frequency	Cumulative Frequency
3	4	4
4	204	208 (= 4 + 204)
5	3	211 (= 208 + 3)
6	84	295 (= 211 + 84)
7	0	295 (= 295 + 0)
8	103	398 (= 295 + 103)

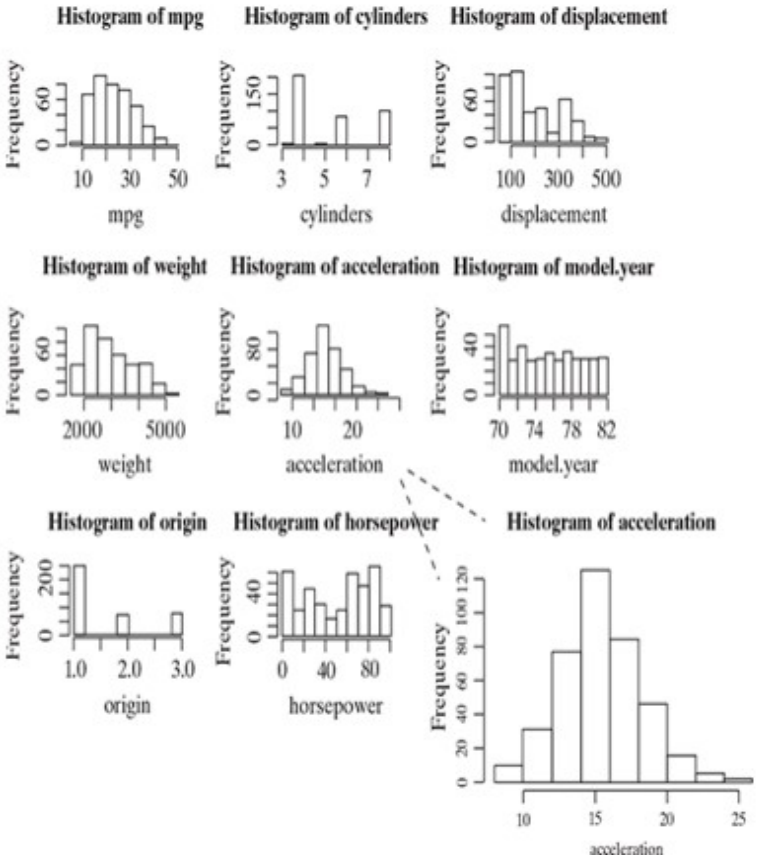
- The frequency is extremely high for data value 4.
- Two other data values where the frequency is quite high are 6 and 8.
- So now if we try to find the quartiles, since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively.
- This way Q1 = 4, median = 4 and Q3 = 8. Since there is no data value beyond 8, there is no upper whisker.
- Also, since both Q1 and median are 4, the band for median falls on the bottom of the box.
- Same way, though the lower whisker could have extended till -2 (Q1 – 1.5 × IQR = 4 – 1.5 × 4 = –2), in reality, there is no data value lower than 3. Hence, the lower whisker is also short. In any case, a value of cylinders less than 1 is not possible.

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.
- Histograms might be of different shapes depending on the nature of the data, e.g. skewness.



- The above figure provides a depiction of different shapes of the histogram that are generally created. These patterns give us a quick understanding of the data and thus act as a great data exploration tool.



- 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

Exploring categorical data

- Statistical measure “mode” is applicable on categorical attributes.
- like mean and median, mode is also a statistical measure for central tendency of a data. Mode of a data is the data value which appears most often.
- In context of categorical attribute, it is the category which has highest number of data values. Since mean and median cannot be applied for categorical variables, mode is the sole measure of central tendency.

For attribute cylinders

Attribute Value	3	4	5	6	8
Count	0.01	0.513	0.008	0.211	0.259

.

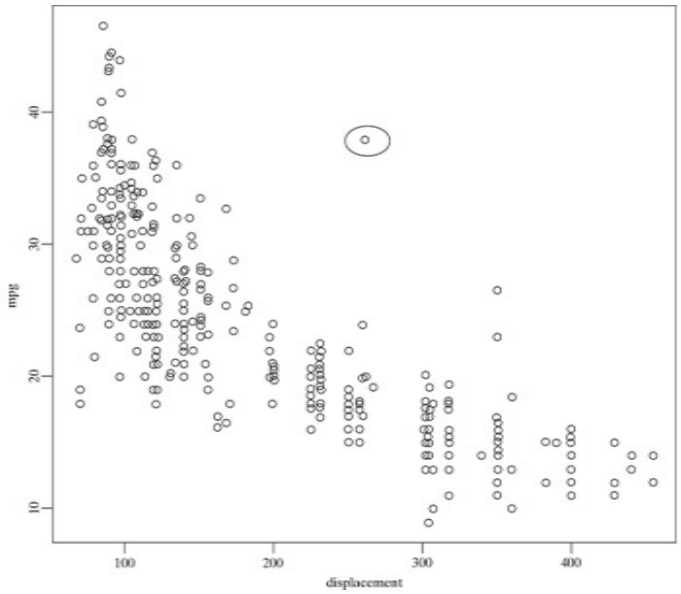
- For cylinders, since the number of categories is less, we can see that the mode is 4, as that is the data value for which frequency is highest.
- More than 50% of data elements belong to the category 4.
- An attribute may have one or more modes.
- Frequency distribution of an attribute having single mode is called ‘unimodal’, two modes are called ‘bimodal’ and multiple modes are called ‘multimodal’.
-

Exploring relationship between variables

- One important angle of data exploration is to explore relationship between attributes.
- There are multiple plots to enable us explore the relationship between variables.
- The basic and most commonly used plot is scatter plot

Scatter plot

- A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables.
- It is a twodimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.
- For example, in a data set there are two attributes – attr_1 and attr_2.
- We want to understand the relationship between two attributes, i.e. with a change in value of one attribute, say attr_1, how does the value of the other attribute, say attr_2, changes.
- We can draw a scatter plot, with attr_1 mapped to x-axis and attr_2 mapped in y-axis.
- So, every point in the plot will have value of attr_1 in the x-coordinate and value of attr_2 in the ycoordinate.
- As in a two-dimensional plot, attr_1 is said to be the independent variable and attr_2 as the dependent variable
- catter plot of ‘displacement’ and ‘mpg’. ‘displacement’ as the x-coordinate and ‘mpg’ as the y-coordinate. The scatter plot comes as in



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.
- For example the attributes ‘cylinders’, ‘model.year’, and ‘origin’ as categorical and we try to examine the variation of one with respect to the other.
- As attribute ‘cylinders’ reflects the number of cylinders in a car and assumes values 3, 4, 5, 6, and 8.
- Attribute ‘model.year’ captures the model year of each of the car and ‘origin’ gives the region of the car, the values for origin 1, 2, and 3 corresponding to North America, Europe, and Asia.

Origin \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
1	22	20	18	29	15	20	22	18	22	23	7	13	20
2	5	4	5	7	6	6	8	4	6	4	9	4	2
3	2	4	5	4	6	4	4	6	8	2	13	12	9

- The above cross tab shows relationship between attributes ‘model. year’ and ‘origin’
- It help us understand the number of vehicles per year in each of the regions North America, Europe, and Asia.
- In another way, we can get the count of vehicles per region over the different years.

DATA QUALITY AND REMEDIATION

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.
- However, it is not realistic to expect that the data will be flawless.

There are at least two types of problems:

1. Certain data elements without a value or data with 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.
- Similarly, if we are trying to predict poll results using a training data which doesn't comprise of a right mix of voters from different segments such as age, sex, ethnic diversities, etc., the prediction is bound to be a failure.
- 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:

- Errors in data collection results 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 either in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067) or in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.).
- 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.

Data remediation

- The issues in data quality, need to be remediated, if the right amount of efficiency has to be achieved in the learning activity.

- Out of the two major areas ,the first one can be remedied by proper sampling technique.
- However, human errors are bound to happen, no matter whatever checks and balances we put in. Hence, proper remedial steps need to be taken for the second area. handle outliers and missing

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 amend those values, you may consider one of the following approaches.
- However, if the outliers are natural, i.e. the value of the data element is surprisingly high or low because of a valid reason, then we should not amend it.

a. **Remove outliers**: If the number of records which are outliers is not many, a simple approach may be to remove them.
b. 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.
c. 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.

Handling missing values

- In a data set, one or more data elements may have missing values in multiple records., it can be caused by omission on part of the surveyor or a person who is collecting sample data or by the responder, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response.
- It may happen that a specific question (based on which the value of a data element originates) is not applicable to a person or object with respect to which data is collected.
- There are multiple strategies to handle missing value of data elements.

Some of those strategies are

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 a 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 because of reduction in the training data size.

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 imputed with the mean, median, or mode of the remaining values under the same attribute.
- For qualitative attributes, all missing values are imputed by the mode of all remaining values of the same attribute.
- However, another strategy may be identify the similar types of observations whose values are known and use the mean/median/mode of those known values.
- If 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.
- The other approach is that we can take a similarity based mean or median.

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.

DATA PRE-PROCESSING

Dimensionality Reduction

- High-dimensional data sets 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 performs 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.
- The principal components are a linear combination of the original variables.

- They are orthogonal to each other. Since principal components are uncorrelated, they capture the maximum amount of variability in the data.
- However, the only challenge is that the original attributes are lost due to the transformation. Another commonly used technique which is used for dimensionality reduction is **Singular Value Decomposition (SVD).**

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 (or contributes almost no information) 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.