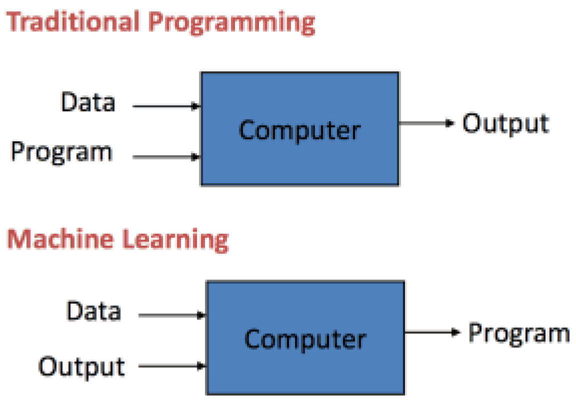
MODULE-1

Machine Learning basics - Learning algorithms - Supervised, Unsupervised, Reinforcement, Overfitting, Underfitting, Hyper parameters and Validation sets, Estimators -Bias and Variance. Challenges in machine learning. Simple Linear Regression, Logistic Regression, Performance measures - Confusion matrix, Accuracy, Precision, Recall, Sensitivity, Specificity, Receiver Operating Characteristic curve(ROC), Area Under Curve(AUC).

1.1 BASICS OF MACHINE LEARNING

Machine Learning is an application of artificial intelligence where a computer/machine learns from the past experiences (input data) and makes future predictions. The performance of such a system should be at least human level.



- Traditional Programming: Data and program is run on the computer to produce the output.
- Machine Learning: Data and output is run on the computer to create a program. This program can be used in traditional programming.

APPLICATIONS OF MACHINE LEARNING

Sample applications of machine learning:

- Web search: ranking page based on what you are most likely to click on.
- Computational biology: rational design drugs in the computer based on past experiments.
- Finance: decide who to send what credit card offers to. Evaluation of risk on credit offers. How to decide where to invest money.
- E-commerce: Predicting customer churn. Whether or not a transaction is fraudulent.
- Space exploration: space probes and radio astronomy.
- Robotics: how to handle uncertainty in new environments. Autonomous. Self-driving car.
- Information extraction: Ask questions over databases across the web.

- Social networks: Data on relationships and preferences. Machine learning to extract value from data.
- Debugging: Use in computer science problems like debugging. Labor intensive process. Could suggest where the bug could be.

KEY ELEMENTS OF MACHINE LEARNING

- There are tens of thousands of machine learning algorithms and hundreds of new algorithms are developed every year.
- Every machine learning algorithm has three components:
    - ✓ Representation: how to represent knowledge. Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.
    - ✓ Evaluation: the way to evaluate candidate programs (hypotheses). Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.
    - ✓ Optimization: the way candidate programs are generated known as the search process. For example combinatorial optimization, convex optimization, constrained optimization

TYPES OF LEARNING

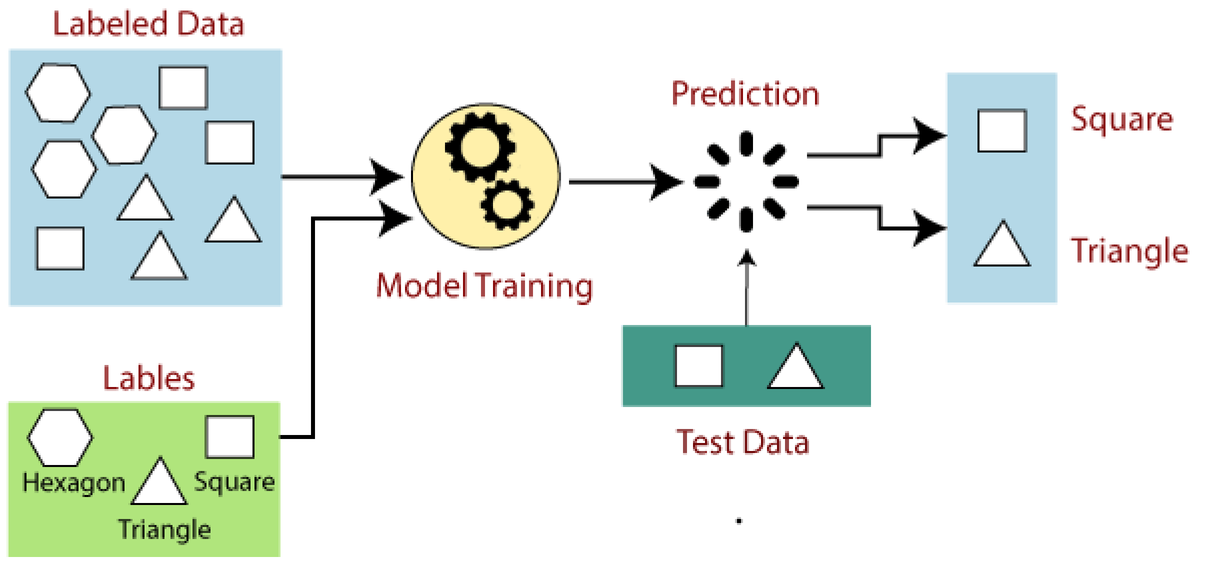
- There are four types of machine learning:
    - ✓ Supervised learning: (also called inductive learning) Training data includes desired outputs. This is spam this is not, learning is supervised.
    - ✓ Unsupervised learning: Training data does not include desired outputs. Example is clustering. It is hard to tell what is good learning and what is not.
    - ✓ Semi-supervised learning: Training data includes a few desired outputs.
    - ✓ Reinforcement learning: Rewards from a sequence of actions. AI types like it, it is the most ambitious type of learning.
- Supervised learning is the most mature, the most studied and the type of learning used by most machine learning algorithms. Learning with supervision is much easier than learning without supervision.
- Inductive Learning is where we are given examples of a function in the form of data (x) and the output of the function (f(x)). The goal of inductive learning is to learn the function for new data (x).
- Classification: when the function being learned is discrete.
- Regression: when the function being learned is continuous.
- Probability Estimation: when the output of the function is a probability.

1.2. SUPERVISED LEARNING

- Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.
- In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
- Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).
- In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

HOW SUPERVISED LEARNING WORKS?

- In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.
- The working of Supervised learning can be easily understood by the below example and diagram:



- Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.
- If the given shape has four sides, and all the sides are equal, then it will be labelled as a Square.

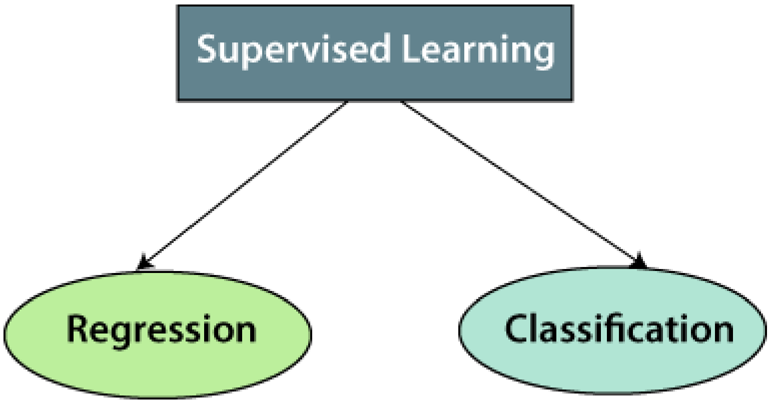
- If the given shape has three sides, then it will be labelled as a triangle.
- If the given shape has six equal sides then it will be labelled as hexagon.
- Now, after training, we test our model using the test set, and the task of the model is to identify the shape.
- The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

STEPS INVOLVED IN SUPERVISED LEARNING:

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training dataset, test dataset, and validation dataset.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

TYPES OF SUPERVISED MACHINE LEARNING ALGORITHMS:

- Supervised learning can be further divided into two types of problems:



1.2.1. Regression

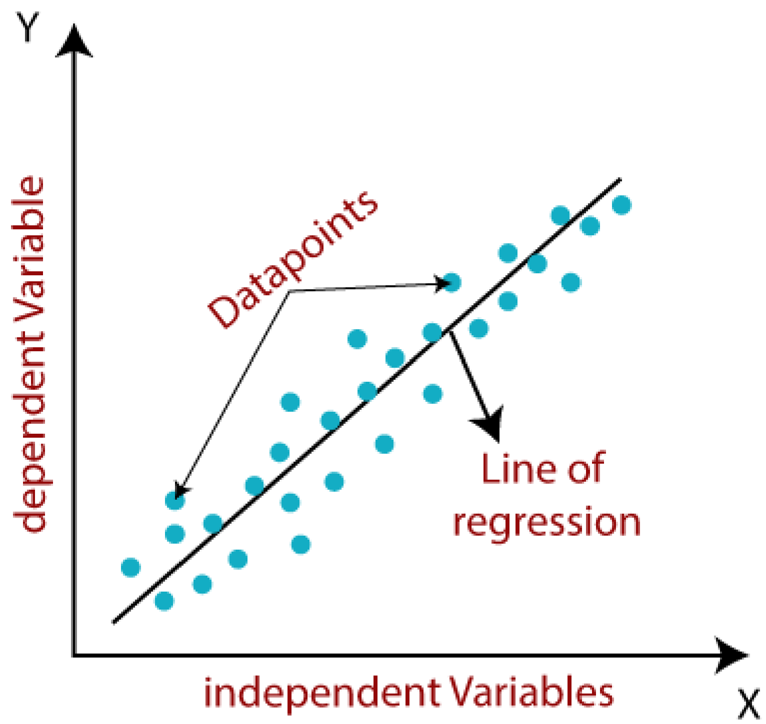
- Regression algorithms are used if there is a relationship between the input variable and the output variable.
- It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.
- Below are some popular Regression algorithms which come under supervised learning:

1.2.1.1. Linear Regression

- ✓ Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

- ✓ Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

- ✓ The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:, y= a+bx+ ε

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a= intercept of the line (Gives an additional degree of freedom)
b = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a regression line.

Cost function

- ✓ The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- ✓ Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- ✓ We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

- ✓ For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\text{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}(y_i \; - (a_1x_i + a_0))^2$$

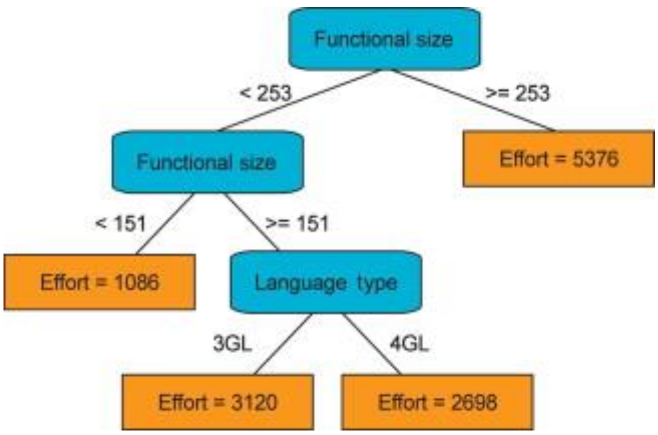
Where,

N=Total number of observation
Yi = Actual value
(a1xi+a0)= Predicted value.

1.2.1.2 . Regression Trees

Regression trees, a variant of decision trees, aim to predict outcomes we consider real numbers — such as the optimal prescription dosage, the cost of gas next year or the number of expected Covid cases this winter.

Regression models attempt to determine the relationship between one dependent variable and a series of independent variables that split off from the initial data set.



1.2.1.3. Non-Linear Regression

Nonlinear regression is a mathematical model that fits an equation to certain data using a generated line. As is the case with a linear regression that uses a straight-line equation (such as Ỵ= c + m x), nonlinear regression shows association using a curve, making it nonlinear in the parameters

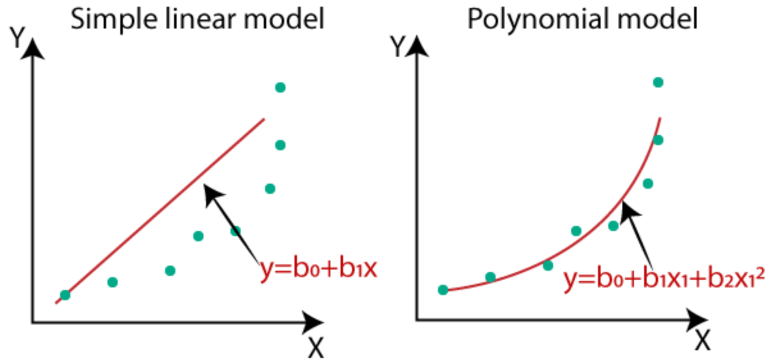
Bayesian Linear Regression

- ✓ In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response, y, is not estimated as a single value, but is assumed to be drawn from a probability distribution.

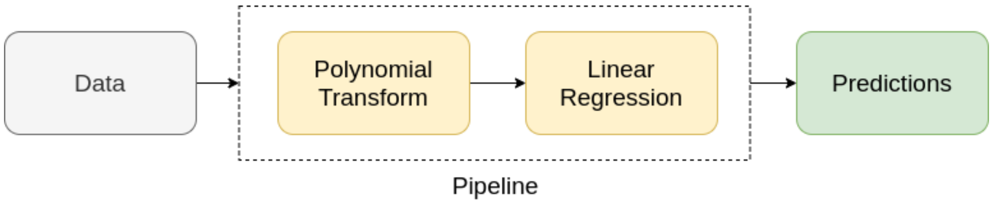
Polynomial Regression

- ✓ Helps identify the curvilinear relationship between independent and dependent variables.
- ✓ Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

y= b0+b1x1+ b2x12+ b2x13+...... bnx1n



- ✓ In the above image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.

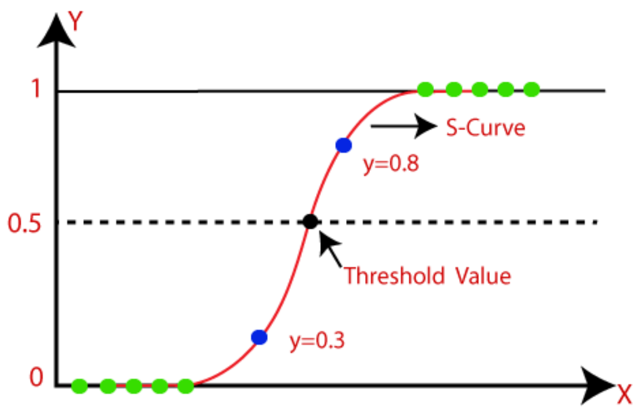


1.2.2. Classification

- 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.

1.2.2.1. Logistic Regression :

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.
- Assumptions for Logistic Regression:
  - ✓ The dependent variable must be categorical in nature.
  - ✓ The independent variable should not have multi-collinearity.
- Equation:

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Type of Logistic Regression:

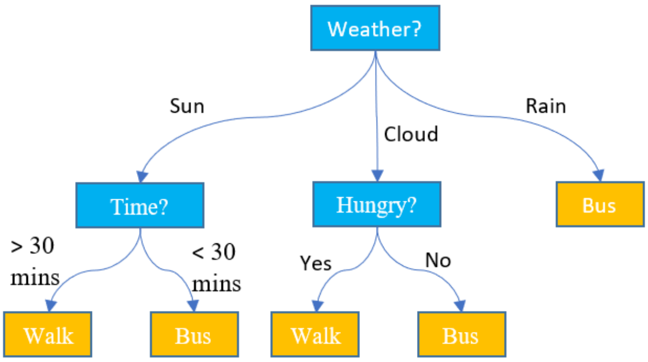
On the basis of the categories, Logistic Regression can be classified into three types:

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

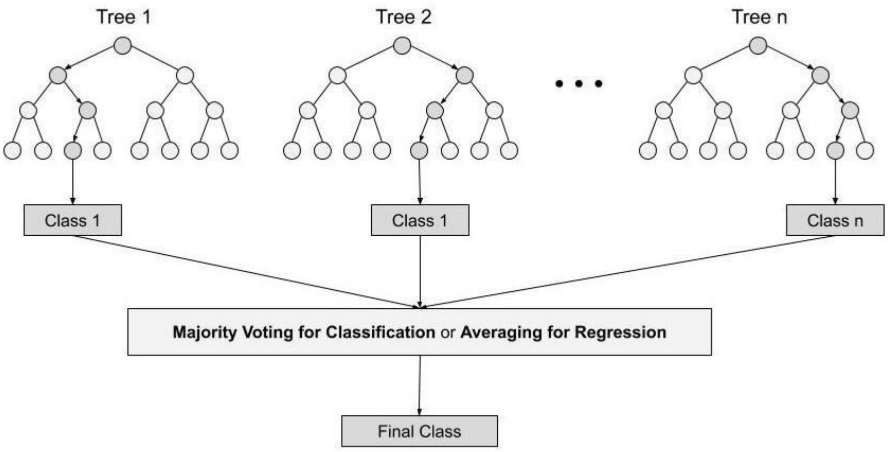
1.2.2.2. Decision Trees

- Decision trees can be used for classification as well as regression problems.
- The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits.
- It starts with a root node and ends with a decision made by leaves.



1.2.2.3. Random Forest

It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.



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.

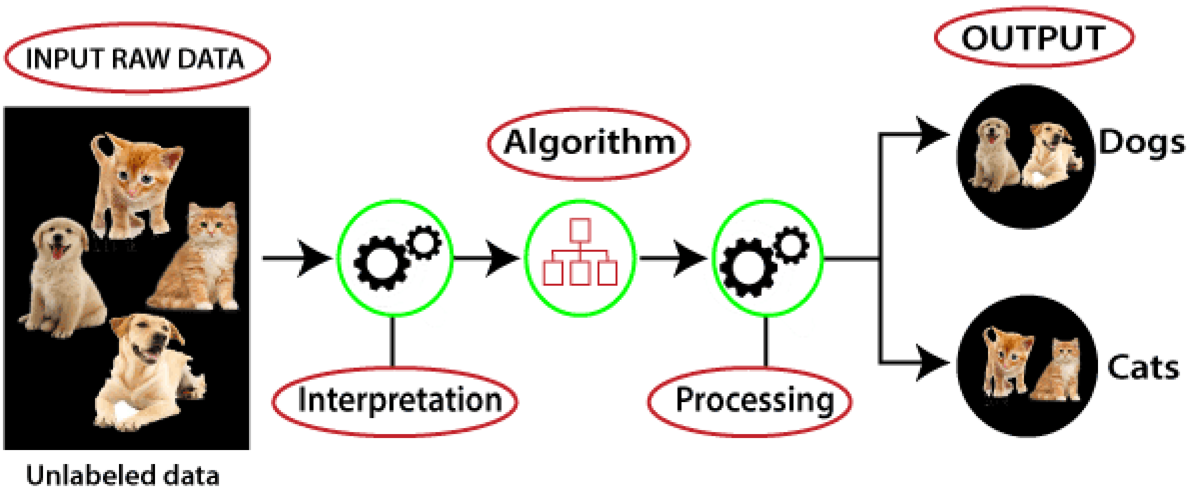
1.3. UNSUPERVISED LEARNING

- 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.
- 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.

Why use 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.

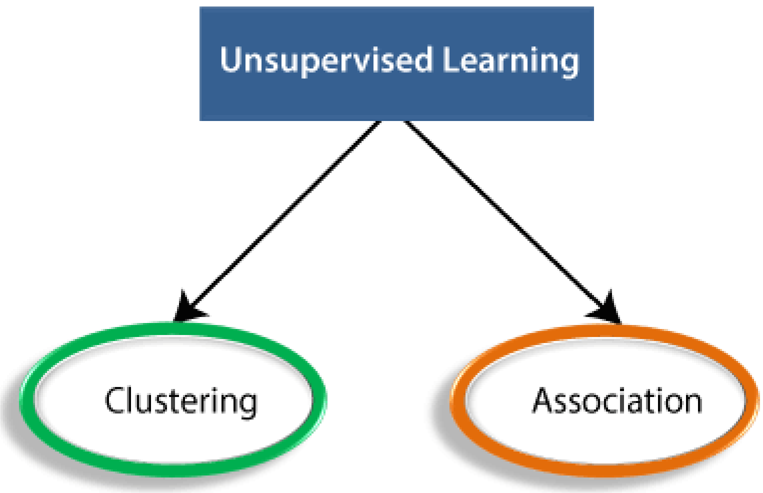
Working of Unsupervised Learning



- Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given.
- Now, this unlabeled input data is fed to the machine learning model in order to train it.
- Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
- Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

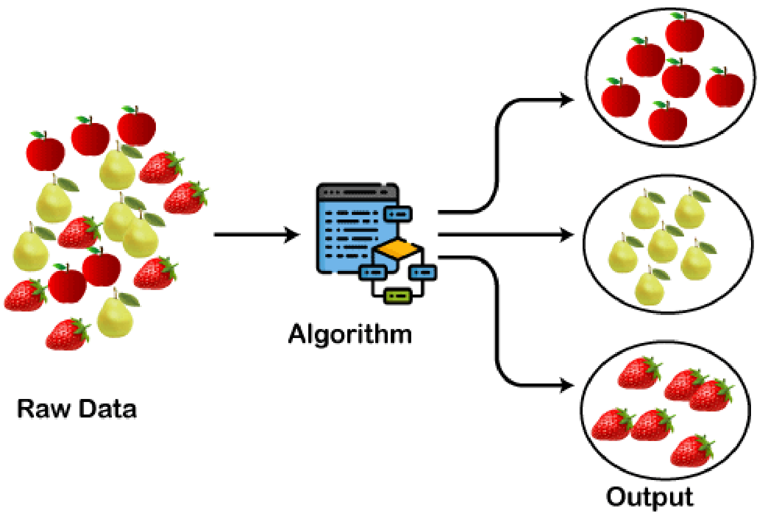
Types of Unsupervised Learning Algorithm:

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.
- For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:

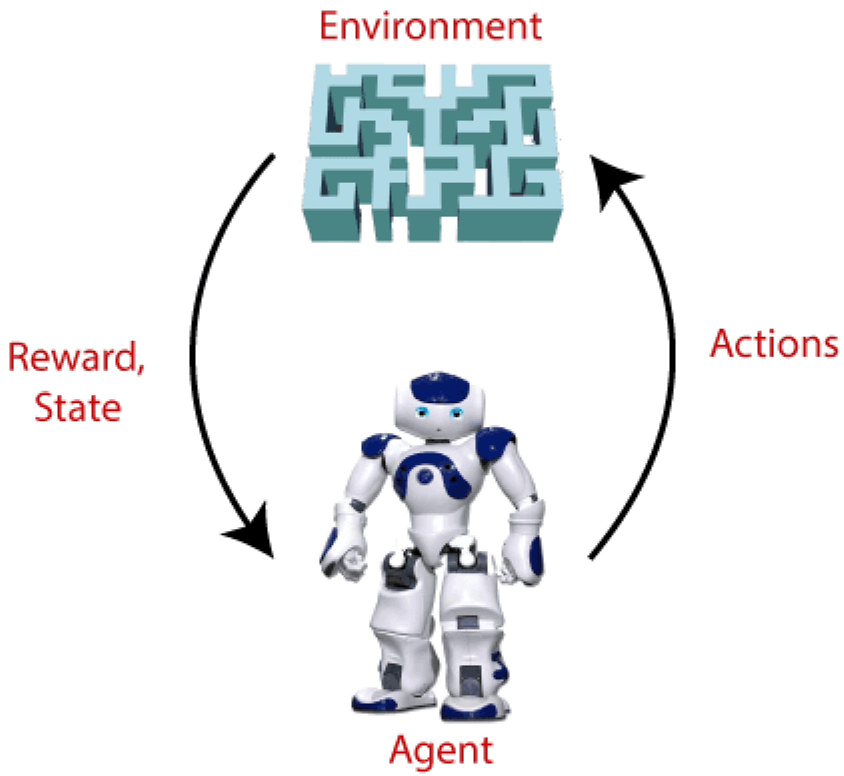
Advantages of Unsupervised Learning

- 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.

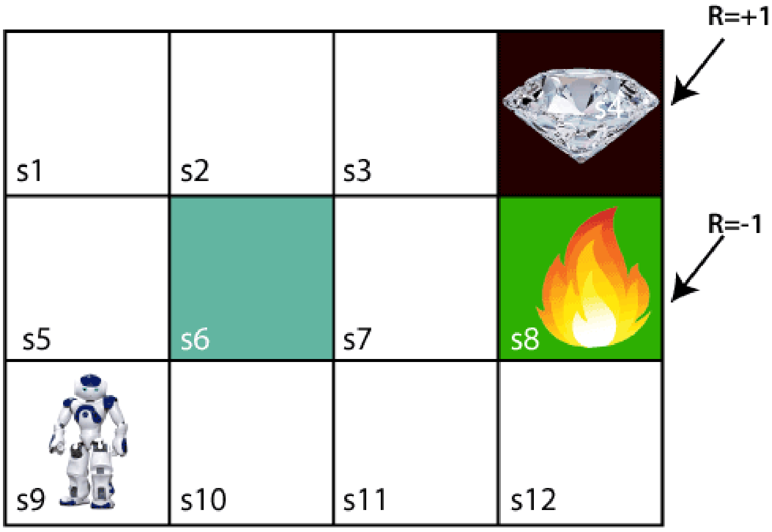
1.4. REINFORCEMENT LEARNING

- 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, unlike supervised learning.
- Since there is no labeled data, so the agent is bound to learn by its experience only.
- RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc.
- The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.
- The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, we can say that "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." How a Robotic dog learns the movement of his arms is an example of Reinforcement learning.
- 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.
- Example: Suppose there is an AI agent present within a maze environment, and his goal is to find the diamond. The agent interacts with the environment by performing some actions, and based on those actions, the state of the agent gets changed, and it also receives a reward or penalty as feedback.
- The agent continues doing these three things (take action, change state/remain in the same state, and get feedback), and by doing these actions, he learns and explores the environment.
- The agent learns that what actions lead to positive feedback or rewards and what actions lead to negative feedback penalty. As a positive reward, the agent gets a positive point, and as a penalty, it gets a negative point.



- Terms used in Reinforcement Learning
  - ✓ Agent(): An entity that can perceive/explore the environment and act upon it.
  - ✓ Environment(): A situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.
  - ✓ Action(): Actions are the moves taken by an agent within the environment.
  - ✓ State(): State is a situation returned by the environment after each action taken by the agent.
  - ✓ Reward(): A feedback returned to the agent from the environment to evaluate the action of the agent.
  - ✓ Policy(): Policy is a strategy applied by the agent for the next action based on the current state.
  - ✓ Value(): It is expected long-term retuned with the discount factor and opposite to the short-term reward.
  - ✓ Q-value(): It is mostly similar to the value, but it takes one additional parameter as a current action (a).
- Key Features of Reinforcement Learning
  - ✓ In RL, the agent is not instructed about the environment and what actions need to be taken.
  - ✓ It is based on the hit and trial process.
  - ✓ The agent takes the next action and changes states according to the feedback of the previous action.
  - ✓ The agent may get a delayed reward.
  - ✓ The environment is stochastic, and the agent needs to explore it to reach to get the maximum positive rewards.
  - ✓ How does Reinforcement Learning Work?
  - ✓ To understand the working process of the RL, we need to consider two main things:
  - ✓ Environment: It can be anything such as a room, maze, football ground, etc.
  - ✓ Agent: An intelligent agent such as AI robot.
  - ✓ Let's take an example of a maze environment that the agent needs to explore. Consider the below image:

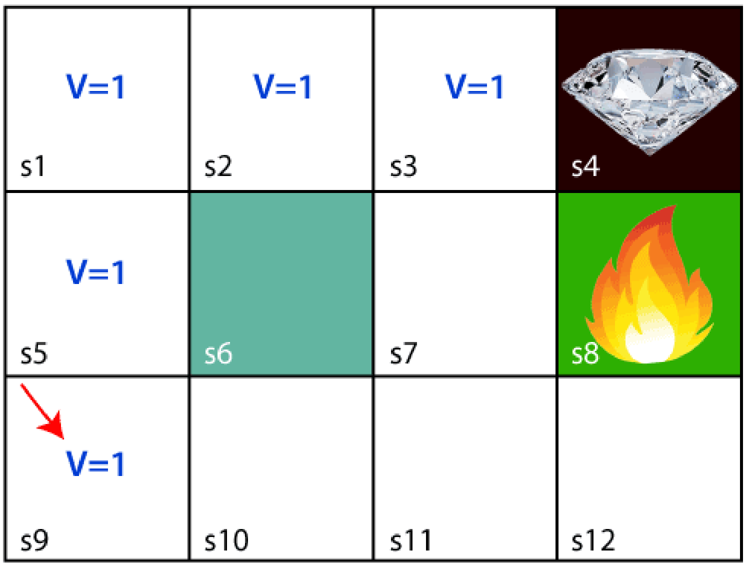


In the above image, the agent is at the very first block of the maze. The maze is consisting of an S6 block, which is a wall, S8 a fire pit, and S4 a diamond block.

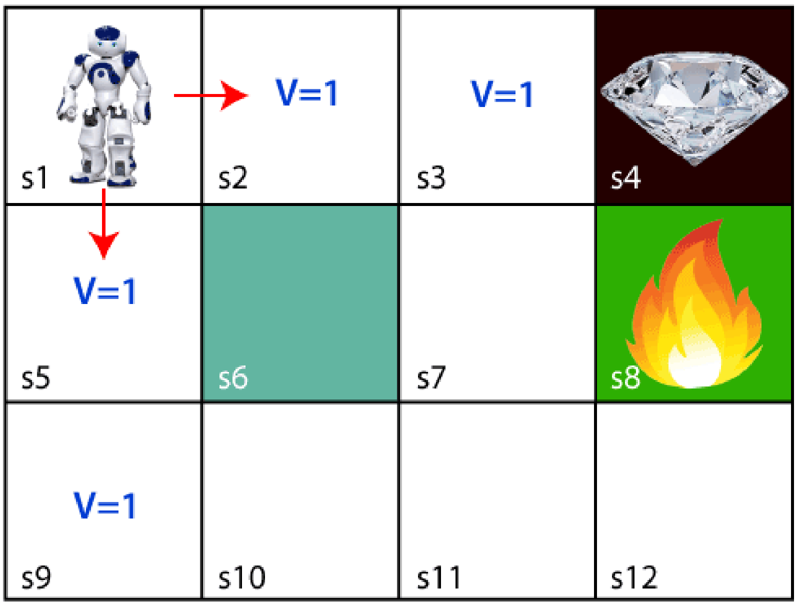
The agent cannot cross the S6 block, as it is a solid wall. If the agent reaches the S4 block, then get the +1 reward; if it reaches the fire pit, then gets -1 reward point. It can take four actions: move up, move down, move left, and move right.

The agent can take any path to reach to the final point, but he needs to make it in possible fewer steps. Suppose the agent considers the path S9-S5-S1-S2-S3, so he will get the +1-reward point.

The agent will try to remember the preceding steps that it has taken to reach the final step. To memorize the steps, it assigns 1 value to each previous step. Consider the below step:



Now, the agent has successfully stored the previous steps assigning the 1 value to each previous block. But what will the agent do if he starts moving from the block, which has 1 value block on both sides? Consider the below diagram:



It will be a difficult condition for the agent whether he should go up or down as each block has the same value. So, the above approach is not suitable for the agent to reach the destination. Hence to solve the problem, we will use the Bellman equation, which is the main concept behind reinforcement learning.

$V(s) = \max [R(s,a) + \gamma V(s`)]$

Where,

$V(s)$= value calculated at a particular point.

$R(s,a)$ = Reward at a particular state s by performing an action.

γ = Discount factor

$V(s`)$ = The value at the previous state.

In the above equation, we are taking the max of the complete values because the agent tries to find the optimal solution always.

So now, using the Bellman equation, we will find value at each state of the given environment. We will start from the block, which is next to the target block.

For 1st block:

$V(s3) = \max [R(s,a) + \gamma V(s`)]$, here $V(s')$= 0 because there is no further state to move.

V(s3)= max[R(s,a)]=> V(s3)= max[1]=> V(s3)= 1.

For 2nd block:

V(s2) = max [R(s,a) + γV(s`)], here γ= 0.9(lets), V(s')= 1, and R(s, a)= 0, because there is no reward at this state.

V(s2)= max[0.9(1)]=> V(s)= max[0.9]=> V(s2) =0.9

For 3rd block:

V(s1) = max [R(s,a) + γV(s`)], here γ= 0.9(lets), V(s')= 0.9, and R(s, a)= 0, because there is no reward at this state also.

V(s1)= max[0.9(0.9)]=> V(s3)= max[0.81]=> V(s1) =0.81

For 4th block:

V(s5) = max [R(s,a) + γV(s`)], here γ= 0.9(lets), V(s')= 0.81, and R(s, a)= 0, because there is no reward at this state also.

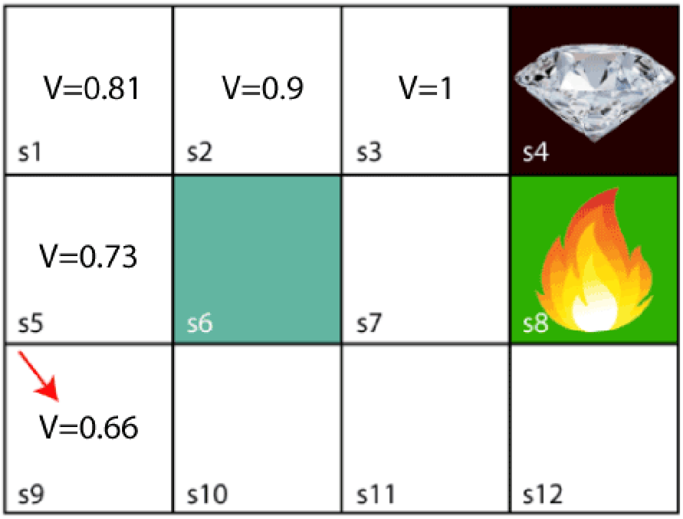
V(s5)= max[0.9(0.81)]=> V(s5)= max[0.81]=> V(s5) =0.73

For 5th block:

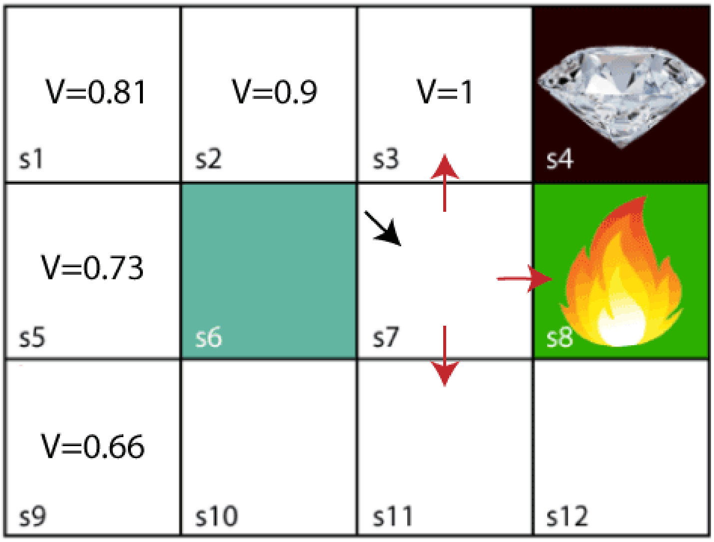
V(s9) = max [R(s,a) + γV(s`)], here γ= 0.9(lets), V(s')= 0.73, and R(s, a)= 0, because there is no reward at this state also.

V(s9)= max[0.9(0.73)]=> V(s4)= max[0.81]=> V(s4) =0.66

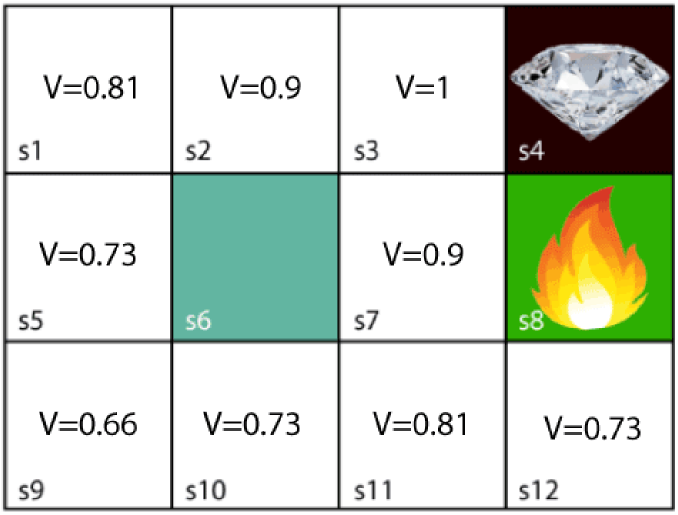
Consider the below image:



Now, we will move further to the 6th block, and here agent may change the route because it always tries to find the optimal path. So now, let's consider from the block next to the fire pit.



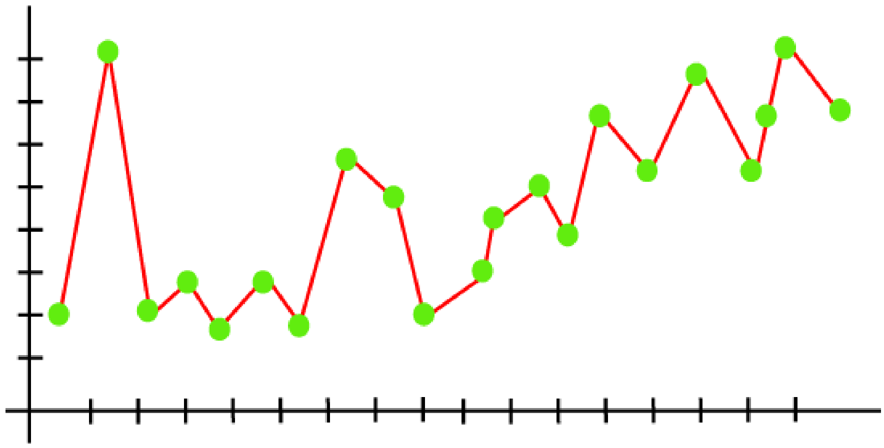
Now, the agent has three options to move; if he moves to the blue box, then he will feel a bump if he moves to the fire pit, then he will get the -1 reward. But here we are taking only positive rewards, so for this, he will move to upwards only. The complete block values will be calculated using this formula. Consider the below image:



1.4 UNDERFITTING AND OVERFITTING

- Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.
- The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.

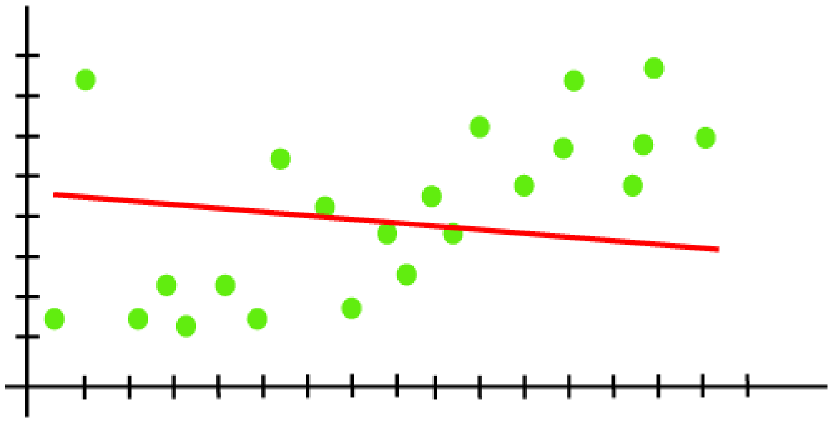
- Overfitting is the main problem that occurs in supervised learning.
- Example: The concept of the overfitting can be understood by the below graph of the linear regression output:



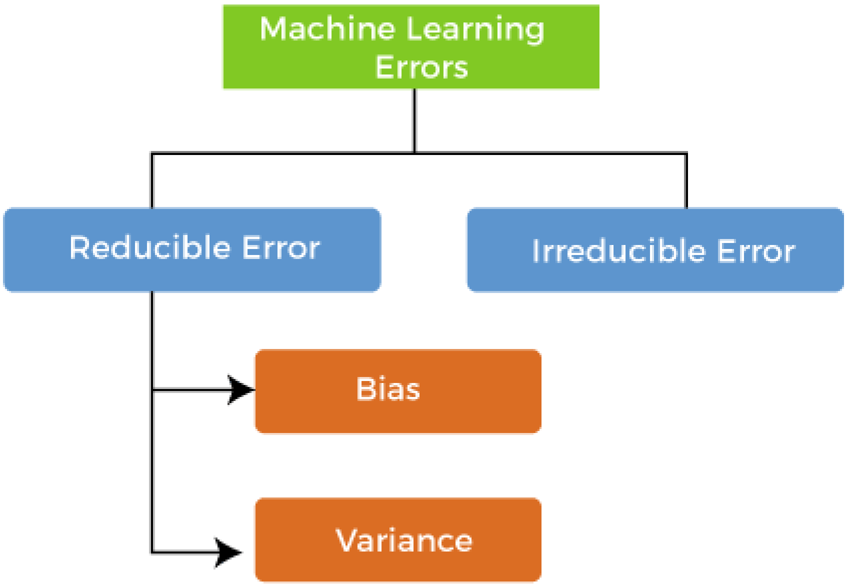
- As we can see from the above graph, the model tries to cover all the data points present in the scatter plot. It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

Underfitting

- Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data
- In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.



1.5. THE ESTIMATORS- BIAS AND VARIANCE



BIAS :

- While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias.
- It can be defined as an inability of machine learning algorithms such as Linear Regression to capture the true relationship between the data points.
- Each algorithm begins with some amount of bias because bias occurs from assumptions in the model, which makes the target function simple to learn.
- A model has either:
- Low Bias: A low bias model will make fewer assumptions about the form of the target function.
- High Bias: A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset. A high bias model also cannot perform well on new data.

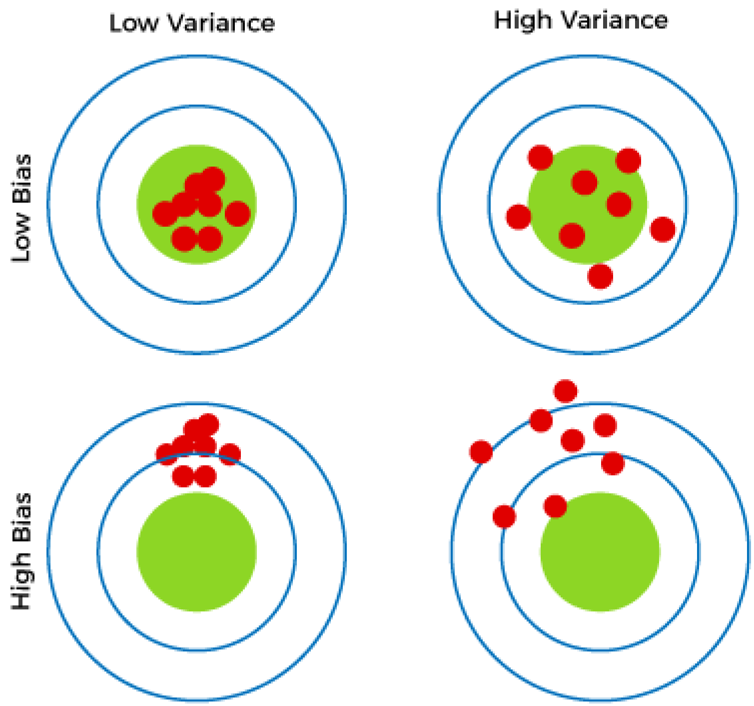
VARIANCE :

- The variance would specify the amount of variation in the prediction if the different training data was used.
- In simple words, variance tells that how much a random variable is different from its expected value.
- Ideally, a model should not vary too much from one training dataset to another, which means the algorithm should be good in understanding the hidden mapping between inputs and output variables.
- Variance errors are either of low variance or high variance.
- Low variance means there is a small variation in the prediction of the target function with changes in the training data set.

- At the same time, High variance shows a large variation in the prediction of the target function with changes in the training dataset.
- A model that shows high variance learns a lot and perform well with the training dataset, and does not generalize well with the unseen dataset.
- As a result, such a model gives good results with the training dataset but shows high error rates on the test dataset.

Different Combinations of Bias-Variance

There are four possible combinations of bias and variances, which are represented by the below diagram:



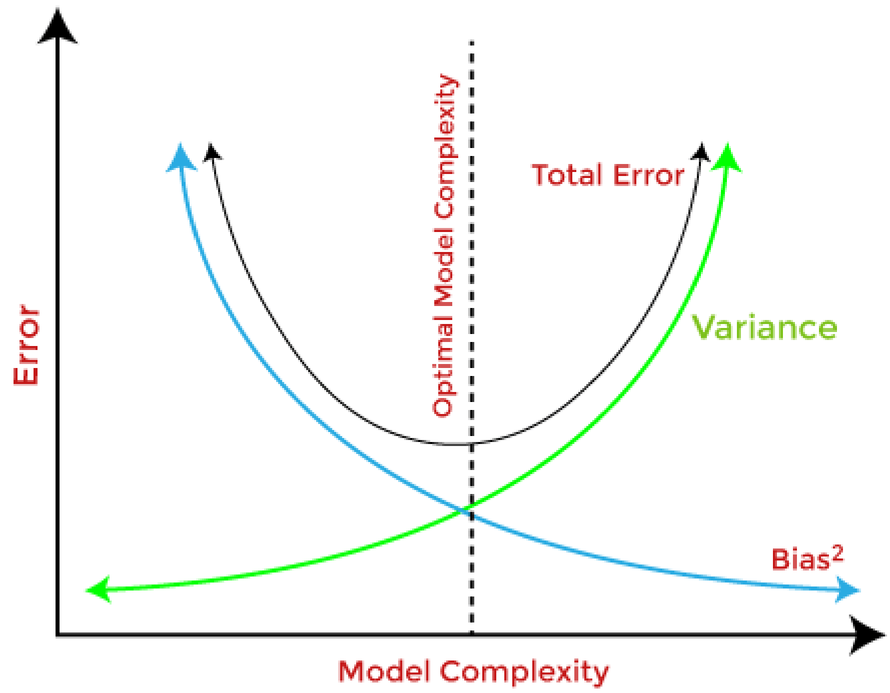
Low-Bias,Low-Variance: The combination of low bias and low variance shows an ideal machine learning model. However, it is not possible practically.

Low-Bias, High-Variance: With low bias and high variance, model predictions are inconsistent and accurate on average. This case occurs when the model learns with a large number of parameters and hence leads to an overfitting

High-Bias, Low-Variance: With High bias and low variance, predictions are consistent but inaccurate on average. This case occurs when a model does not learn well with the training dataset or uses few numbers of the parameter. It leads to underfitting problems in the model.

High-Bias,High-Variance:With high bias and high variance, predictions are inconsistent and also inaccurate on average.

Bias-Variance Trade-Off



- For an accurate prediction of the model, algorithms need a low variance and low bias. But this is not possible because bias and variance are related to each other:
- If we decrease the variance, it will increase the bias.
- If we decrease the bias, it will increase the variance.
- Bias-Variance trade-off is a central issue in supervised learning.
- Ideally, we need a model that accurately captures the regularities in training data and simultaneously generalizes well with the unseen dataset.
- Unfortunately, doing this is not possible simultaneously.
- Because a high variance algorithm may perform well with training data, but it may lead to overfitting to noisy data.
- Whereas, high bias algorithm generates a much simple model that may not even capture important regularities in the data.
- So, we need to find a sweet spot between bias and variance to make an optimal model.
- Hence, the Bias-Variance trade-off is about finding the sweet spot to make a balance between bias and variance errors.

1.6. PERFORMANCE MEASURES

Confusion Matrix

- A confusion matrix is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
- A confusion matrix is a table that categorizes predictions according to whether they match the actual value

		Predicted class: *P*	Predicted class: *N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
Actual Class	*N*	False Positives (FP)	True Negatives (TN)

		Prediction: Cat	Prediction: Dog
Actual	Cat	15	35
Actual	Dog	40	10

- Classification Rate or Accuracy is given by the relation:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Recall (Sensitivity) : the ratio of the total number of correctly classified positive examples divide by the total number of positive examples.

- High Recall indicates the class is correctly recognized

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

- Precision: total number of correctly classified positive examples by the total number of predicted positive examples.

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

- High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

- Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP).

Eg1:

Suppose a computer program for recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats. Of the eight dogs identified, five actually are dogs while the rest are cats. Compute the precision and recall of the computer program.

TP = 5

FP = 8-5 = 3

FN = 12-5 = 7

The precision P is $P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{5}{5+3} = \frac{5}{8}$

The recall R is $R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{5}{5+7} = \frac{5}{12}$

Eg 2:

Let there be 10 balls (6 white and 4 red balls) in a box and let it be required to pick up the red balls from them. Suppose we pick up 7 balls as the red balls of which only 2 are actually red balls. What are the values of precision and recall in picking red ball?

TP = 2

FP = 7-2 = 5

FN = 4-2 = 2

The precision P is $P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{2}{2+5} = \frac{2}{7}$

The recall R is $R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{2}{2+2} = \frac{1}{2}$

Eg 3:

Assume the following: A database contains 80 records on a particular topic of which 55 are relevant to a certain investigation. A search was conducted on that topic and 50 records were retrieved. Of the 50 records retrieved, 40 were relevant. Construct the confusion matrix for the search and calculate the precision and recall scores for the search.

Actual \ Predicted	Relevant	Irrelevant
Relevant	40	15
Irrelevant	10	15

TP = 40

FP = 10

FN = 15

The precision P is: 40/(40+10) = 4/5

The recall R is : 40/(40+15) = 8/11

1. $\text{Accuracy} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
2. $\text{Error rate} = 1 - \text{Accuracy}$
3. $\text{Sensitivity} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}$
4. $\text{Specificity} = \dfrac{\text{TN}}{\text{TN} + \text{FP}}$
5. $F\text{-measure} = \dfrac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$

Receiver Operating Characteristic (ROC)

- ROC stands for Receiver Operating Characteristic
- The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields.

- They are now increasingly used in machine learning and data mining research.

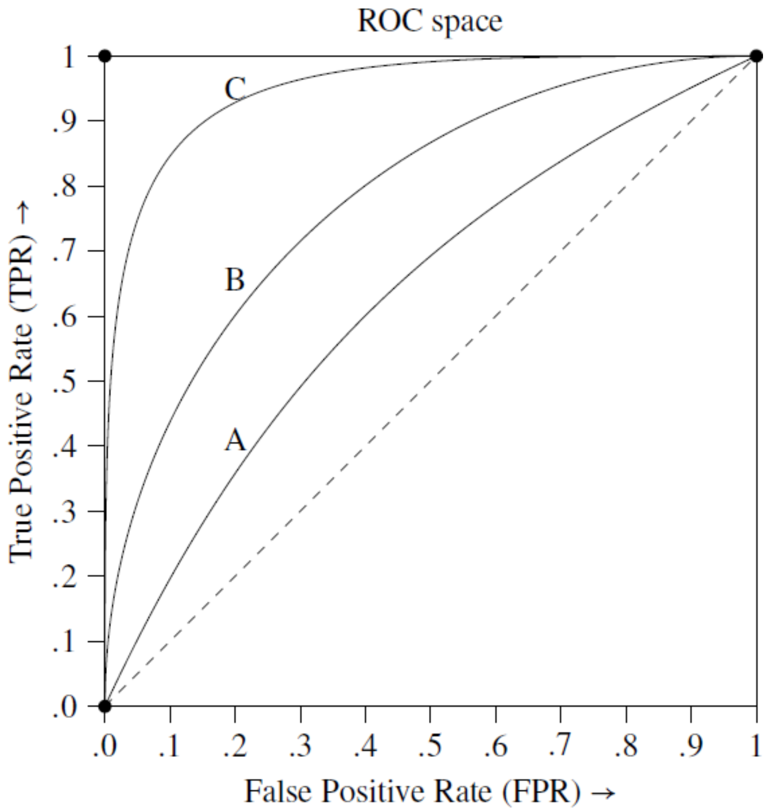
- ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones.

- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

$$\text{TPR} = \text{True Positive Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \text{False Positive Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- In the case of certain classification algorithms, the classifier may depend on a parameter.
- Different values of the parameter will give different classifiers and these in turn give different values to TPR and FPR.
- The ROC curve is the curve obtained by plotting in the ROC space the points (TPR ; FPR) obtained by assigning all possible values to the parameter in the classifier.



•The closer the ROC curve is to the top left corner (0; 1) of the ROC space, the better the accuracy of the classifier.

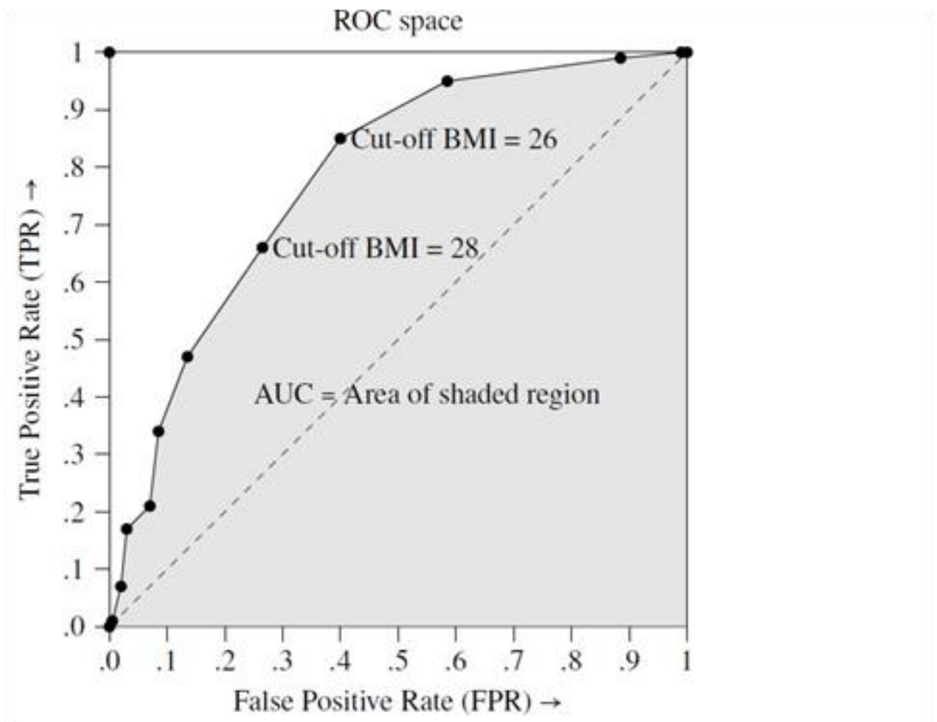
•Among the three classifiers A, B, C with ROC, the classifier C is closest to the top left corner of the ROC space.

•Hence, among the three, C gives the best accuracy in predictions.

Eg:

Cut-off value of BMI	Breast cancer		Normal persons		TPR	FPR
	TP	FN	FP	TN		
18	100	0	200	0	1.00	1.000
20	100	0	198	2	1.00	0.990
22	99	1	177	23	0.99	0.885
24	95	5	117	83	0.95	0.585
26	85	15	80	120	0.85	0.400
28	66	34	53	147	0.66	0.265
30	47	53	27	173	0.47	0.135
32	34	66	17	183	0.34	0.085
34	21	79	14	186	0.21	0.070
36	17	83	6	194	0.17	0.030
38	7	93	4	196	0.07	0.020
40	1	99	1	199	0.01	0.005

Data on breast cancer for various values of BMI



ROC curve of data in showing the points closest to the perfect prediction point

<u>Area Under curve(AUC)</u>

- The measure of the area under the ROC curve is denoted by the acronym AUC.
- The value of AUC is a measure of the performance of a classifier.
- For the perfect classifier, AUC = 1.0.