Name: Dhaval kumar Vijay kumar Patel

MSc CS

TA-2 Assignment

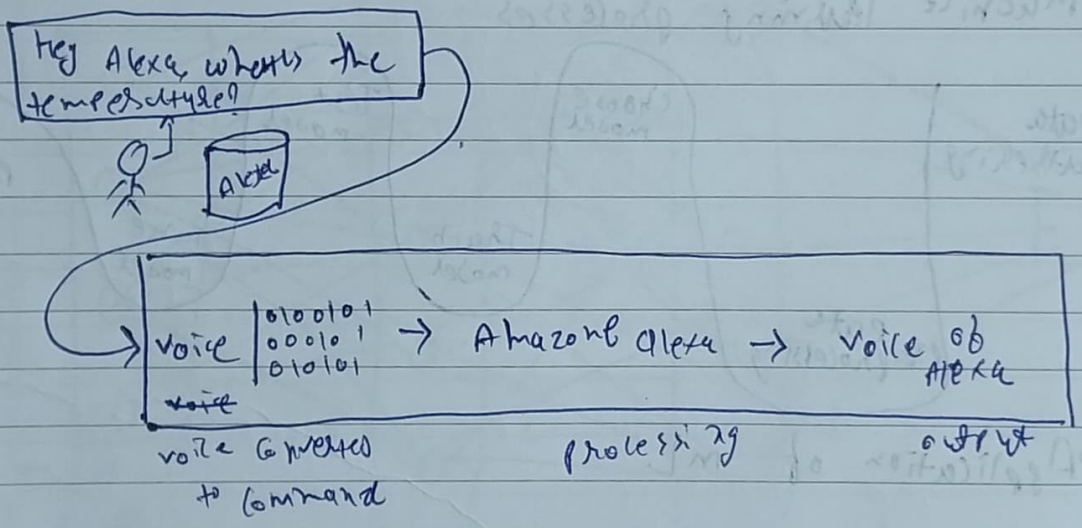
Date of submission 02/01/2022

Artificial Intelligence

q.1) Illustrate with an example the difference between machine learning, deep learning and Artificial Intelligence.

→ Artificial Intelligence.

Artificial Intelligence (AI) is the process of imparting data, information and human intelligence to machine. The main goal of AI is to develop self-reliant machine that can think and act like human. These machines can mimic human behavior and perform tasks by learning and problem-solving. Most of the AI systems simulate natural intelligence to solve complex problems. Basically Artificial Intelligence means Ability of a machine to imitate intelligent human behavior.



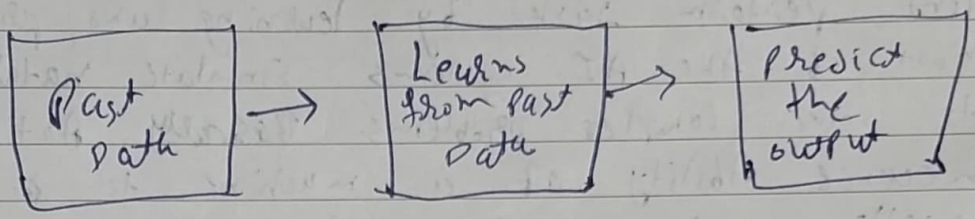
↳ Application of AI

- machine Translate
- self Driving vehicals
- AI Robot
- Speech recognition applications like Apple's siri or ok google

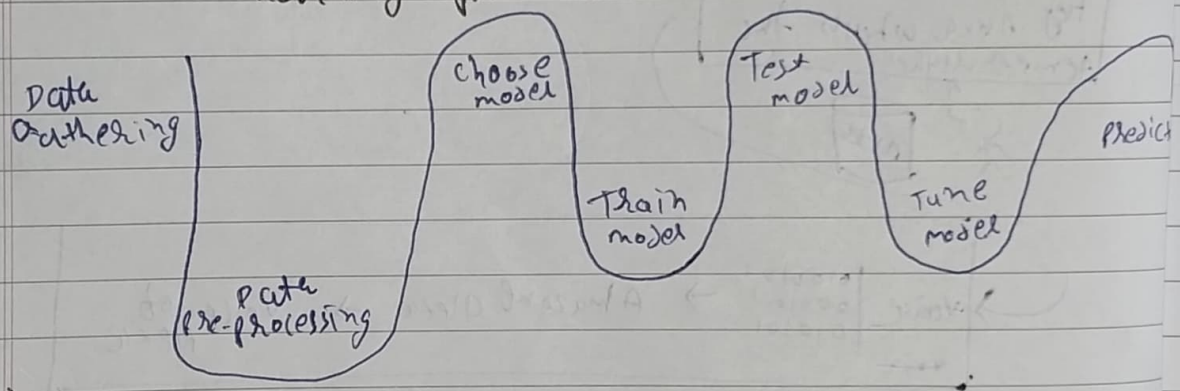
Machine Learning

Machine learning (ML) is a discipline of computer science that uses computer algorithms and analytics to build predictive models that can solve business problems.

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 task in T, as measured by P, improves with experience E.



machine learning processes



Application of ML

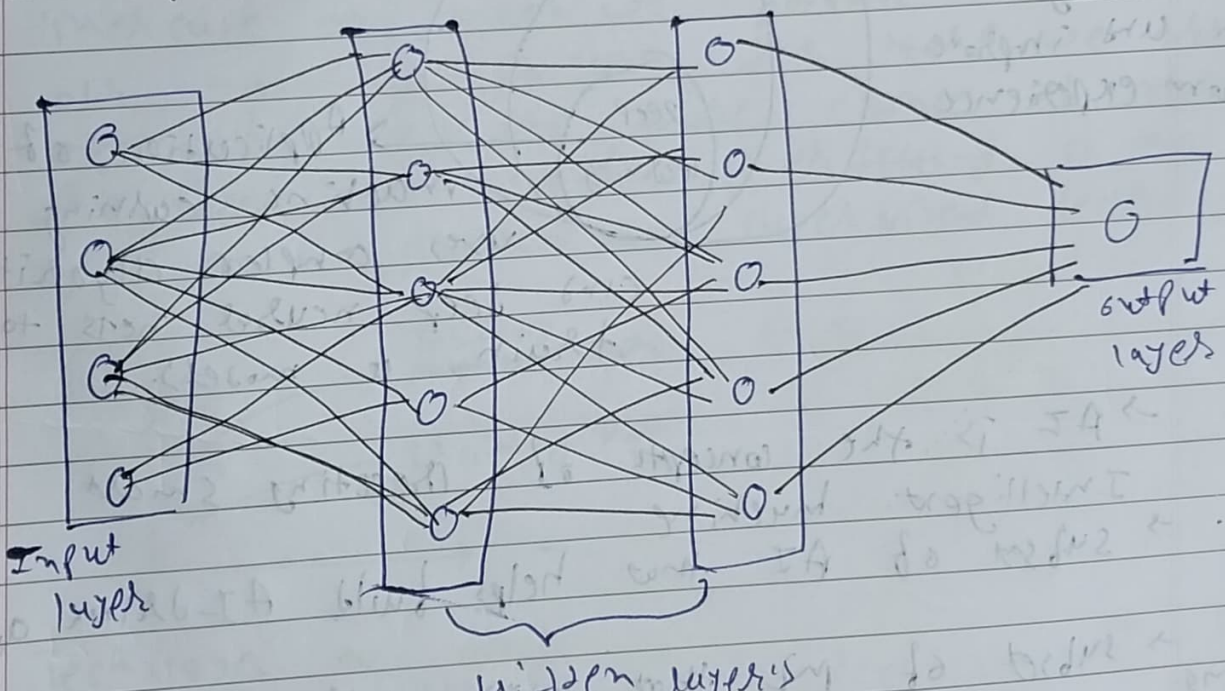
- sales forecasting for different product
- fraud analysis in banking
- Product recommendation
- Stock price prediction

↳ Deep Learning

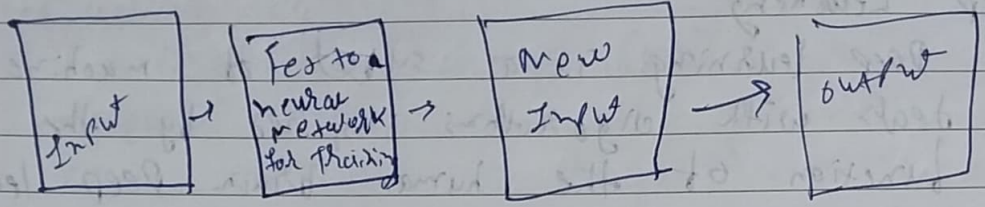
Deep learning is a subset of machine learning that deals with algorithms inspired by the structure and function of the human brain. Deep learning algorithms can work with an enormous amount of both structured and unstructured data. Deep learning's core concept lies in artificial neural networks, which enable machine to make decisions.

The major difference between deep learning vs machine learning is the way data is presented to the machine. ML algorithms usually require structured data, whereas deep learning networks works on multiple layers of artificial neural networks

↳ Simple neural network

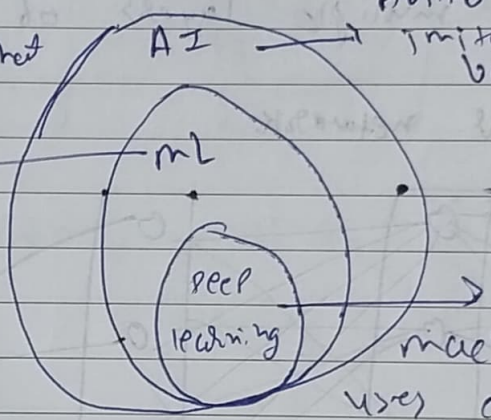


How It process



↳ Application of Deep learning

- cancer tumor detection
- captionbot for captioning an image
- music generating
- Image & coloring
- object detection



Application of AI that Allows a system to automatically learn and improve from experience

Ability of a machine to imitate intelligent human behavior

Application of machine learning that uses complex algorithms and deep neural nets to train a models.

AI → AI is the concept of creating smart Intelligent machine

ML → subset of AI that helps build AI-driven applicati

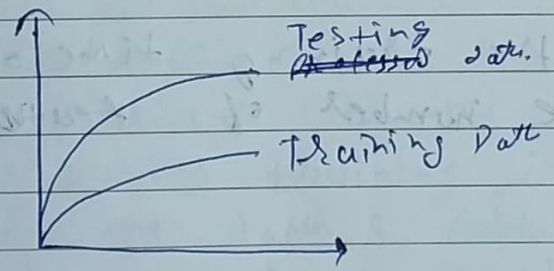
Deep learning → subset of ML that uses vast volumes of data and complex algorithms to train a model

Q.2 Explain what is overfitting and underfitting. Explain the techniques to avoid overfitting and underfitting

→ overfitting

overfitting occurs when our machine learning model tries to cover all data point 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 overfitting model has low bias and high variance.

The chance of occurence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chance of couring the overfitted model. overfitting is the main problem in supervised learning



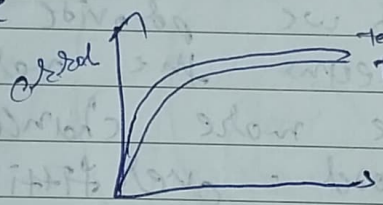
Techniques to avoid

- Cross-validation
- Training with more Data
- Removing features
- Early stopping the training
- Regularization
- Ensembing

↳ Underfitting

- Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduce the accuracy and produces unreliable predictions. Underfitting has high bias and low variance



Technique to avoid

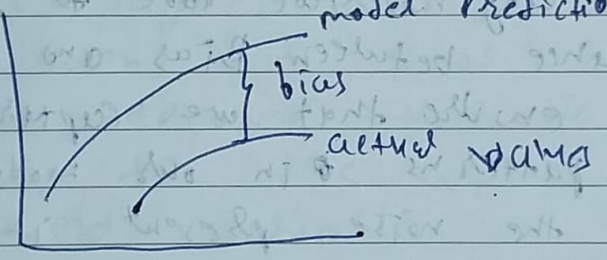
- Increasing the training time of model
- increasing the number of features.

Q.3 Explain the terms bias and variance with respect to machine Learning. Discuss the bias-variance trade-off in machine learning

↳ Bias

To make predictions, our model will analyze our data and find patterns in it. Using these pattern we can make generalizations about certain instances in our data. our model after training learn these patterns and applies them to the test ~~to~~ set to predict them.

Bais is the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be predict new data



↳ variance

variance is the very opposite of ~~Bais~~ Bias during training, it allows our model to see the data a certain number of times to find patterns in it. if it does not work on the data for long enough, it will not find pattern and bias occurs. on the other hand, if our model is allowed to view the data too many times, it will learn very well for only the Data. It will capture most patterns

in the data, but it will also learn from the unnecessary data present, or from the noise.

We can define variance as the model's sensitivity to fluctuations in the data. Our model learn from noise. This will cause our model to consider trivial feature as important.

The variance:

The variability of model prediction for a given data point which tell us spreads of our data is called the variance of the model.

↳ Bias-variance trade off

For any model, we have to find the perfect balance between Bias and variance. This just ensure that we capture the essential patterns in our model while ignoring the noise present in it. This is called Bias-variance Tradeoff. It helps optimize the error in our model and keeps it as low as possible.

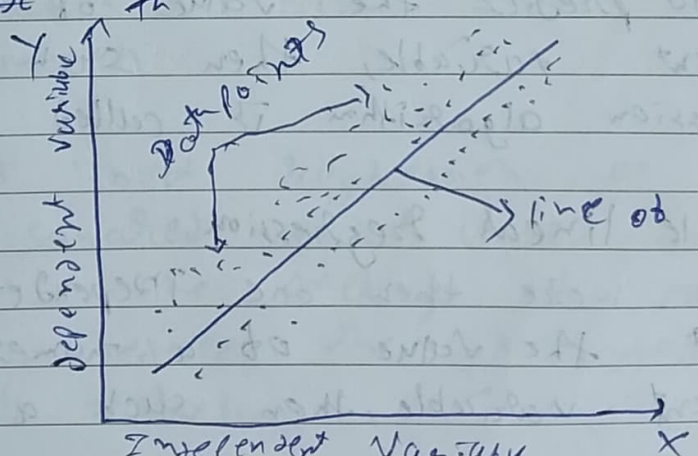
~~Ex~~ ~~Discuss~~

Q.4 Discuss in detail the application of linear regression for prediction

Linear regression is one of the most easiest and most popular machine learning algorithms. It is a statistical method that is used for predictive analysis. Linear ~~Reg &~~ regression makes predictions for continuous/real or numeric variable such as Sales, Salary, age, Product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the 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 slope straight line representing the ~~bet~~ relationship between the variables. ~~Consider~~ It



we can represent a linear regression as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

here,

y = ~~Dependent~~ Dependent variable (Target variable)

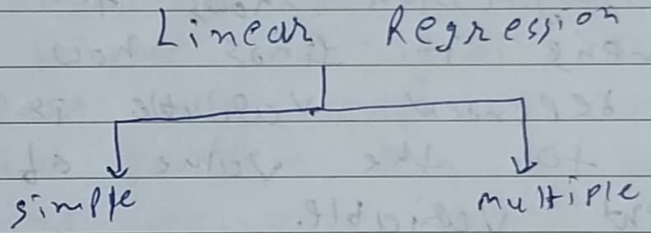
x = Independent variable (predictor variable)

β_0 = Intercept of line
(Given as additional degree of freedom)

β_1 = Linear regression coefficient
(scale factor to each input value)

ε = random error

The value for x and y variable are training datasets ~~of~~ for linear Regression model representation



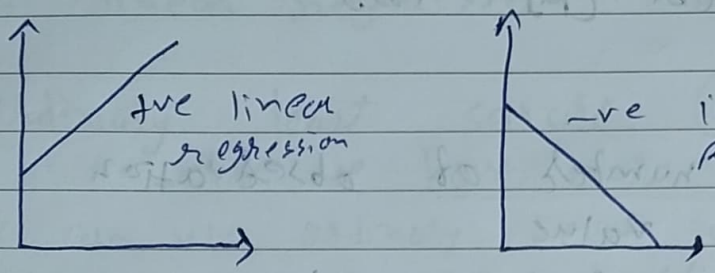
↳ Simple linear regression

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.

↳ Multiple linear Regression

if more then one independent variable to predict the value of a numerical dependent - variable, then such a linear

Regression algorithm is called multiple linear Regression



Cost function:-

- The different values for weights or coefficient of line (b_0, b_1) given 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 occured between the predicted values and actual values.

It can be written as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - (\beta_1 x_i + \beta_0))^2$$

where,

N = Total number of observation
y_i = actual value
$(\beta_1 x_i + \beta_0)$ = predictive value.

Residuals:-

The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter point are close to the line, then the residual will be small and hence the cost function will small.

Q.5 Briefly describe any two techniques used to compute the parameters of the linear regression model

→ (i) Ordinary least square (OLS)

- when we have more than one input we can use ordinary least squares to estimate the values of the coefficients. The OLS procedure seeks to minimize the sum of the square ~~reziv~~ residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the square errors (SSE) together This is the quality that OLS seeks to minimize

$$SSE = \frac{\sum_{i=1}^{n} d_i^2}{n}$$

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations. It is unusual to implement the OLS procedure yourself. Unless as an exercise in linear algebra. It is more likely that you will call a procedure in a linear algebra

library. This procedure is very fast to calculate

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

where,

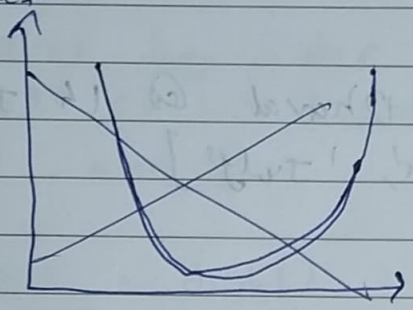
x_i = independent variables
$\bar{x}$ = average of independent variable
y = dependent variables
$\bar{y}$ = average of dependent variable

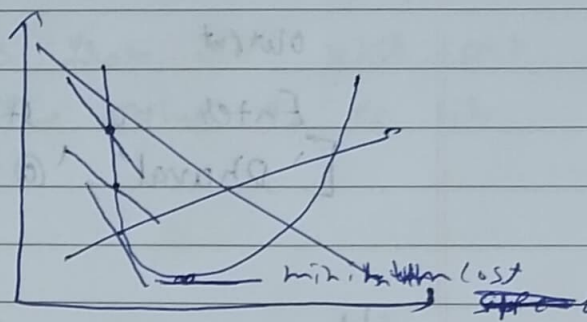
(ii) Gradient Descent

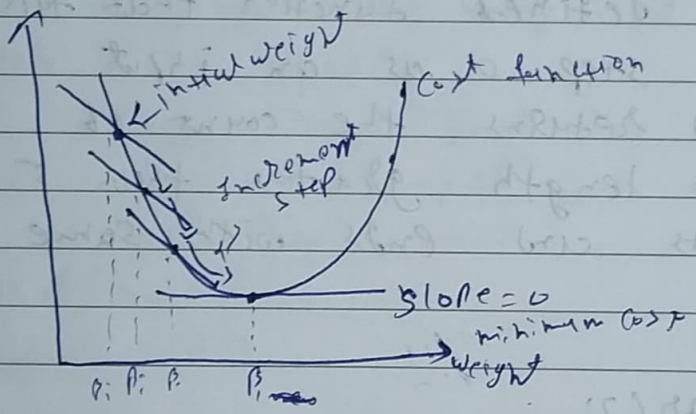
Gradient Descent is known as one of the most commonly used optimization algorithm to train machine learning models by means of minimizing errors between actual and expected results. Further, gradient descent is also used to train neural networks.

In mathematical terminology, optimization algorithm refers to task of minimizing/maximizing an objective function $f(x)$ parameterized by x. Similarly, in ML, optimization is the task of minimizing the cost function parameterized by the

model's parameters. The main objective of gradient descent is to minimize the convex function using iteration of parameter updates. once these machine learning models are optimized, these models can be used as powerful tools for AI and various computer science applications. It helps in finding the local minimum of a function







$$J(\beta_0, \beta_1) = \frac{\sum (\hat{y}_i - y_i)^2}{N}$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\beta_i = \beta_i - \eta \frac{\partial J(\beta_0, \beta_1)}{\partial \beta_i}$$

Q6 write a python program that does/has the following:

a) Takes string inputs from the user and store in a list

```
Str = list("Enter a string: ").split(" ")
Print(Str)
```

Output

Enter a string: Dharval @ 14 July
['Dharval', '@', '14', 'July']

b) A user defined function that takes the list in step. a as an input parameter and returns the count of strings of length grater than 5 that starts and ends with same character.

```
def inputstr():
    lst = list(input("Enter your string:").split(" "))
    print(lst)
    count = 0
    for x in lst:
        if len(x) > 5:
            if x[0] == x[len(x)-1]:
                count = count + 1
    lst.clear()
    return count
Inputstr()
```

Q7

output:

Enter your string: hdahbd jhaduj jsdhs kale Kndiauk ld ud(dncna nmsdcjn

4

Q.7 Write a python program that does/has the following

a.) Takes key value inputs from the user and store in a dictionary. The value is a list of numbers.

```
mydisk = dict()
n= int(input("How many input you want:"))
for x in range(n):
        Key = (input("Enter Key:"))
        Value = int(input("Enter value in numeric form:"))
        mydisk [key] = value
print(mydisk)
mydisc.clear()
```

Output

How many input you want: 3
Enter Key : K1
Enter value in numeric form: 1
Enter Key : K2
Enter value in numeric form: 2
Enter Key K3
Enter value in numeric form: 3

{'K1': 1, 'K2': 2, 'K3': 3}

b.) A user defined function that takes the dictionary in step 4 as an input parameter and return the key that has minimum number of unique elements in its list

2.8 A box contain 3 blue marble, 4 red, 6 green marble and 2 yellow marbles. If three marbles are picked at random, what is the probability that they are all blue?

→ Given that there are

3 blue marble
4 red marble
6 green marble
2 yellow marble

Total marble = 15

Probability that they all are blue is

$$P(\text{all are blue}) = \frac{{}^3C_3}{{}^{15}C_3}$$

$$= \frac{3!}{\frac{15 \times 14 \times 13}{3!}}$$

$$= \frac{1}{455}$$

Q.9 A Bag contains 6 blue balls and 4 Red Balls. 3 balls are picked at random. what is the probability that none of them is Red?

⇒ Given that there are

6 Blue Balls
4 Red Balls
} total ball = 10

→ Probability of getting Red 3 times

$$P(\text{getting Red}) = \frac{{}^4C_3}{{}^{10}C_3}$$

$$= \frac{4 \times 3 \times 2}{10 \times 9 \times 8}$$

$$= \frac{1}{30}$$

there we can say none of them is Red is equal all use blue

∴ P(getting none of them is Red)

$$= P(\text{getting blue}) = \frac{{}^6C_3}{{}^{10}C_3}$$

$$= \frac{6 \times 5 \times 4}{10 \times 9 \times 8}$$

$$= \frac{1}{6}$$

Q.10 The probability that A speaks truth is 3/5 and that of B speaking truth is 4/7. what is probability that they agree in stating the same fact

→ Given $P(A) = \frac{3}{5}$ speaking Truth

$P(B) = \frac{4}{7}$ Speaking Truth

Find probability that A and are agree to the same fact

→ since A and B are independent,

$\therefore P(A \cap B)$

$= P(A) \times P(B)$ [Speaking of both Truth]

$= \frac{3}{5} \times \frac{4}{7}$

$= \frac{12}{35}$

→ Probability both speaking false

$= [1 - P(A)] \times [1 - P(B)]$

$= [1 - \frac{3}{5}] \times [1 - \frac{4}{7}]$

$= \frac{2}{5} \times \frac{3}{7}$

$= \frac{6}{35}$

Q. 12

∴ Probability both agree the same fact

= P(speaking truth) + P(speaking false)

$= \frac{12}{35} + \frac{6}{35}$

$= \frac{12+6}{35}$

$= \frac{18}{35}$

Q.12 Explain in detail any five applications of machine learning

1) virtual personal Assistant:

We have various virtual Assistants such as Google assistant, Alexa, Cortana, Siri, As the name suggestes, they help us in finding the information using our voice instruction. These assistans can help us in various ways just by our voice instructions such as play music call someone, open an email, sheduling an appointment etc.

These assistant record our voice instruction, send it over the server on a cloud, and decode it using ML algorithms and act accordingly

2) Transportation and commuting (cabs) (Uber)

Online cabs booking using ML to an extent. It provides a personalized application which is unique to you. Automatically detects your location and provides options to either go home or office or any other frequent places based on your History and patterns.

It uses ML algorithm layered on top of Historic Trip data to make a more accurate ETA prediction. With the implementation of ML, They saw a 26% accuracy in Delivery and pickup

3) Self driving cars. (TESLA)

ML plays a very important role in self driving cars and we all heard about TESLA. The leader in this business and their current AI is driven by hardware manufacturer NVIDIA, which is based on unsupervised Learning. Algorithm

4) Traffic Alerts (maps)

Google maps we use when we go out some new place and it show us which route may have traffic which is more faster. But how does it know that?

It's a combination of people currently using the service, Historic Data of that route collected over time and few tricks acquired from other companies. Everyone using maps is providing their location, average speed, the route in which they are travelling which in turn help Google collect massive data about traffic, which makes them predict the upcoming traffic and adjust your route according to it.

5) Email spam and malware Filtering:

whenever we recive a new mail, it is filtered automatically as important, normal, and spam we always recive an important mail in our inbox with

the important symbol and spam emails in our spam box and the technology behind this is machine learning. Below are some spam filters used by gmail:

- content filter
- Header filter
- General blacklist filter
- Rules-based filters
- Permission filter

Some ML algorithm such as multi-layer perceptron, Decision tree and Naive Bayes classifier are used for email spam filtering and malware detection

Q.13 Differentiate between Pearsons correlation and spearmans correlation. Give examples illustrating the use of both the measures.

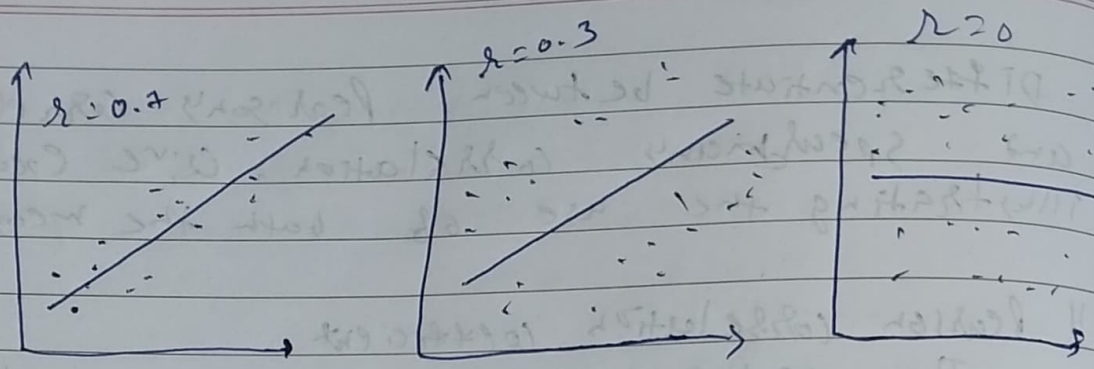
→ ① Pearson correlation coefficient.

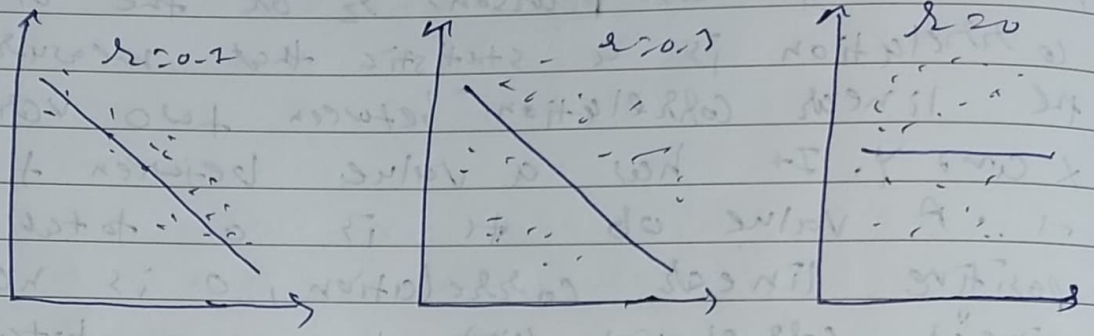
The pearson correlation coefficient also referred to as pearsons r or the bivariate correlation is a statistic that measures the linear correlation between two variables x and y. It has a value between +1 and -1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation.

The pearson correlation can evaluate only a linear relationship between two continuous variable (A relationship is linear only when a change in one variable is associate with a proportional chang in the other variable

Example: we can use the pearson correlation to evaluate whether an increase in age leads to an increase in blood pressure

↳ Below is an example of how the pearson correlation coefficient (r) varies with the strength and the direction of the relationship between the two variables. Note that when no linear relationship could be established, the pearson coefficient yields a value of zero.





↳ Spearman correlation coefficient

Spearman's rank correlation coefficient or spearman's ρ, named after charles spearman is a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables.) It assesses how well the relationship between two variable can be described using a monotonic function

The spearman correlation can evaluate a monotonic relationship between two variables- continue or ordinal and it is based on the ranked values for each variable rather than the raw data

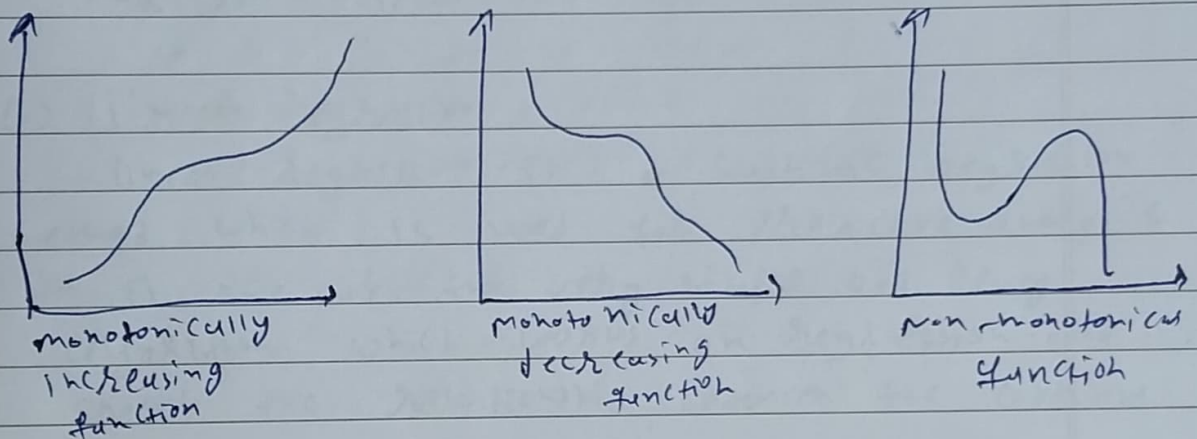
What is monotonic relationship?

- A monotonic relationship is a relationship that does one of the following:

(i)- as the value of one variable increases, so does the value of the other variable, or,

(ii) - as the value of one variable increases, the other variable value decreases.

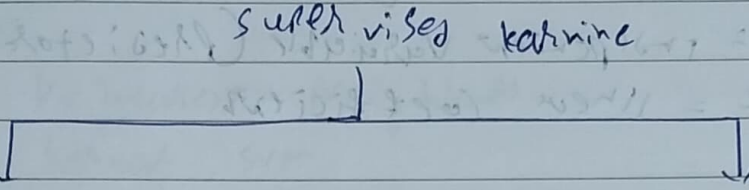
But, not exactly at a constant rate whereas in a linear relationship the rate of increase/decrease is constant



Example:- Whether the order in which employee complete a test exercise is related to the number of month they have been employed or correlation between the IQ of a person with the number of hours spent in front of TV per week.

4. Explain in brief any five supervised learning models.

Supervised learning can be devided into two types of problems:



- → Linear Regression
- → non-linear regression
- → Bayesian linear regression
- → Polynomial Regression
- → Regression Trees

- → Random forest
- → Decision trees
- → Logistic Regression
- → Support Vector machines

(i) linear Regression

- linear Regression is a statistic regression method which is used for predictive analysis
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continue variable
- It is used for solving the regression problem in machine learning
- linear regression show the linear relationship between the independent variable (x) and dependent variable (y), hence called linear regression
- If there is only one input variable (x) then such linear regression called simple linear Reg Regression And if there

If more then one input variable, then its called multiple linear regression

- equation of linear regression

$$Y = mx + c$$

Here,

Y = dependent variable (target variable)

X = independent variable (predictor variable)

m & c = linear coefficient

(ii) Non-linear regression

- non-linear regression is a form of regression analysis in which observational data are ~~model~~ modeled by a function which is a non-linear combination of the model perameters and depends on one or more independent variables. The data are fited by a method of successive approximations.

A simple non-linear regression model is expressed as follows.

$$Y = f(x, \beta) + \epsilon$$

where

X = vector of p predictor

β = vector of k parameter

f(-) = known regression function

ϵ = error term

Alternatively, the model can also be written as

$$Y_i = h[x_i^{(1)}, x_i^{(2)} \ldots x_i^{(m)}; \theta_1, \theta_2, \ldots, \theta_p] + \epsilon_i$$

Where,

Y = Responsive variable

h = function

x = input

θ = parameter to be estimated

Example

k-nearest Neighbours

kernal svm

Naive Bayes

Decision Tree ~~class~~

(iii) Naive Bayes algorithm

- Naive Bayes algorithm is a supervised ~~algo~~ learning algorithm, which is bayes on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training data set.

Bayes theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

(iv) Logistic Regression

- Logistic Regression is another supervised learning algorithm which is used to solve the classification problem. In classification problem, we have dependent variable in a binary or discrete format such as 0 or 1.

- logistic regression algorithm which work on the concept of probability.
- Logistic regression is a type of reg ression, but it is different from the linear regression algorithm in the term how they are used.

- logistic regression uses sigmoid function or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression. The function can be represented as:

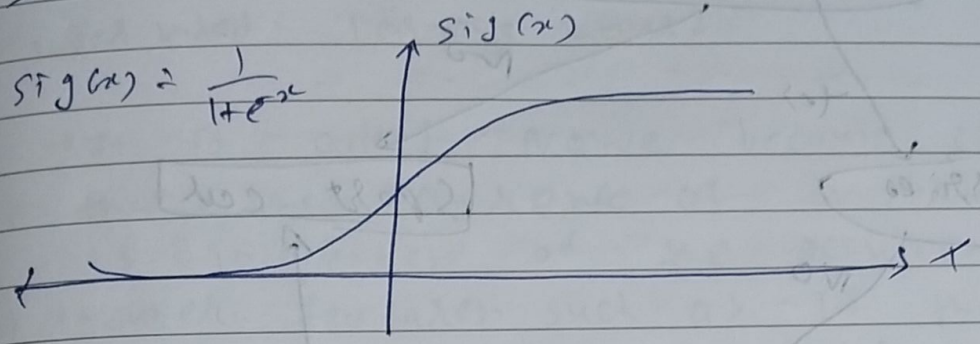
$$f(x) = \frac{1}{1+e^{-x}}$$

where

$f(x)$ = output between the 0 and 1.
x = input to the function
e = base of natural logarithm

when we provide the input values to the function, it gives the

S- curve as follows:



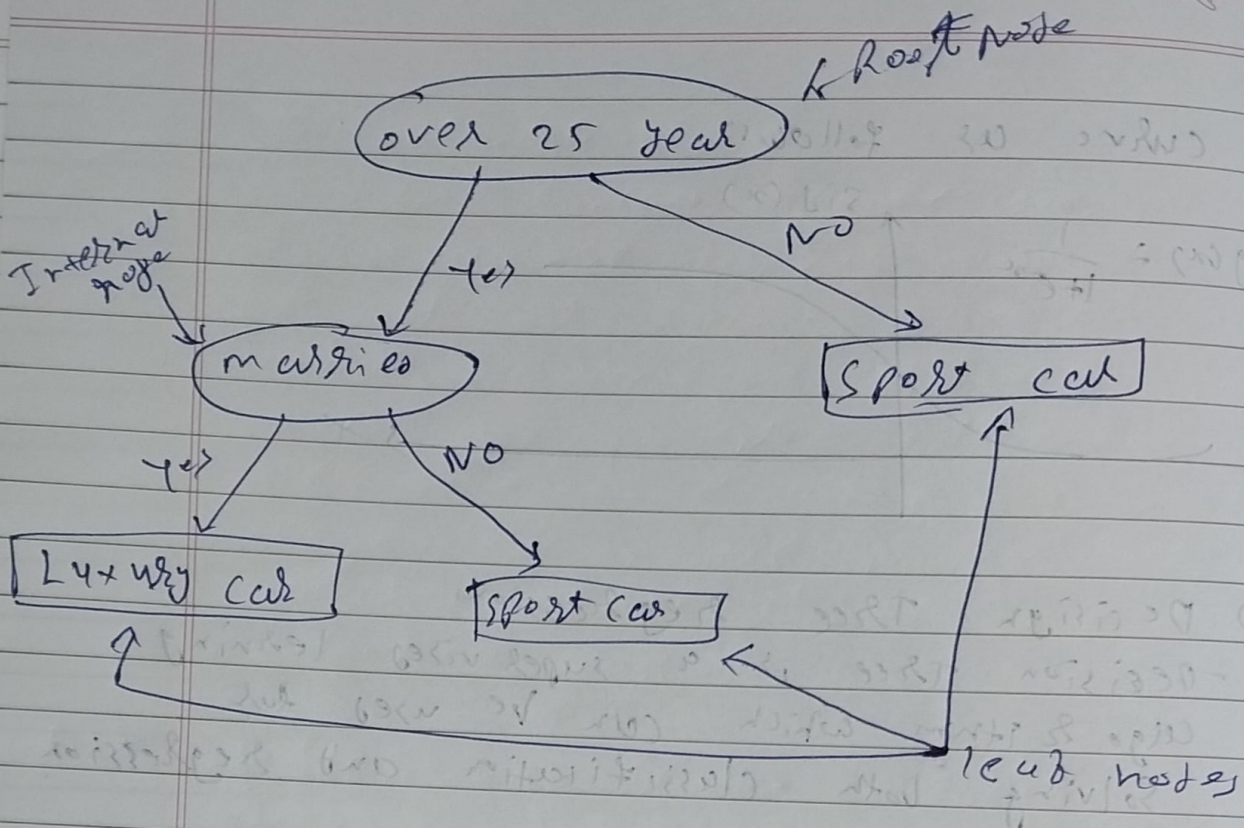
(v) Decision Tree regression:

- Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problem

- It can solve problems for both categorical and numerical data

- Decision Tree regression builds a tree like structure in which each internal node represents the "test" for an attribute each branch represent the result of the test, and each leaf node represents the final decision or result

- A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subset of dataset). These child nodes are further divided into their child nodes, and themselves become the parent node of those nodes. Consider the following


Root Node
over 25 year
Yes
NO
Interest in sports
married
Yes
NO
Sport car
Luxury car
Sport car
leaf nodes

Q.15 what is Naive in Naive Bayes supervised learning model?

→ It is called Naive because it assumes that the occurence of a certain feature is independent of the occurence of other features. such as if the fruits is identified on the base of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individuality contribute to identify that it is an apple without depending on each other.

Q.16 Explain the terms loss function and cost function. Differentiate between both.

loss function:

The loss function is a method of evaluating how well your machine learning algorithm models your featured dataset. In other words loss function are a measurement of how good your model is in term of predicting the expected outcome.

Cost function:

Cost function measure the performance of a machine learning model for a data set cost function quantifies the error between predicted and expected values and presents that error in the form of a single real number. Depending on the problem cost function can be formed in many different ways.

Q.17 Differentiate between parametric model and non-parametric model

~~Parametric met~~

- In case of parametric models, the assumption related to the function form is made and linear model is considered. In case of non parametric model the assumption about the functional form is not made.

- Parametric model are much easier to fit than non-parametric models because parametric ~~mod~~ machine learning models only require the estimation of a set of parameter as the model is identified prior as linear model. In case of non-parametric model one needs to

estimate some arbitrary function which is a much difficult task.

- Parametric models often do not match the unknown function we are trying to estimate. The model performance is comparatively lower than non-parametric model. The estimates done by the parametric models will be further from being true.

- Parametric models are interpretable unlike the non parametric model. This essentially means that one can go for parametric models when the goals is to find inference. Instead, one can go for non parametric model when the goal is to make prediction with higher accuracy and interpretability of inference is not the key q>K.

Q.18 Define the cost function for logistic regression model?

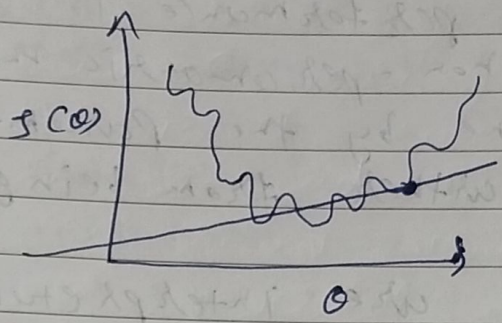
→ In the case of liner Regression, the cost function is:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left[h_\theta (x^{(i)} - y^{(i)} \right]^2$$

But for logistic Regression,

$$h_\theta(x) = g(\theta^T x)$$

It will result in a non-convex cost function. But this result in cost function with local optima's which is a very Big problem for Gradient descent to compute the global optima.

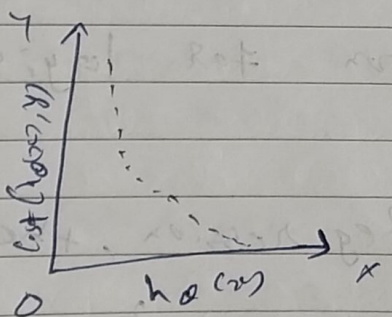


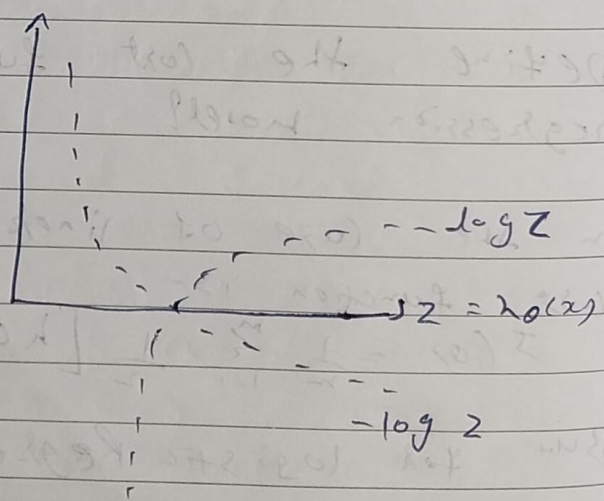
Non convex graph for cost function

So, for logistic regression the cost function is

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

If $y = 1$





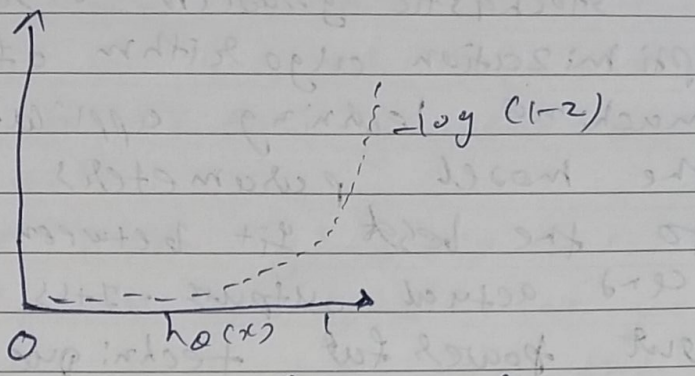
cost = 0 if y = 1, $h_\theta(x) = 1$

But as

$h_\theta(x) \to 0$

cost → Infinity

If y = 0



So,

$$\text{cost}(h_\theta(x), y) = \begin{cases} 0 & \text{if } h_\theta(x) = y \\ \infty & \text{if } y = 0 \text{ and } h_\theta(x) \to 1 \\ \infty & \text{if } y = 1 \text{ and } h_\theta(x) \to 0 \end{cases}$$

$$\text{cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{cost}(h_\theta(x), y)$$

To fit the parameter θ, J(θ) has to be minimized and for the gradient descent is required.

Gradient Descent - looks similar to that of linear Regression but the difference lies in the hypothesis $h_\theta(x)$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Q.19 with respect to gradient descent define the following terms

→ a) Stochastic gradient descent:-

Stochastic gradient descent is an optimization algorithm often used in machine learning applications to find the model parameters that corresponds to the best fit between predicted and actual outputs. It is an inexact but powerful technique. Stochastic gradient descent is widely used in machine learning applications.

b) Batch gradient descent:-

In Batch gradient descent, all the training data is ~~taken~~ taken into consideration to take a single step. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters so that's just one step of gradient descent in one epoch.

c) mini batch gradient descent →

We use a batch of a fixed number of training example which is less than the actual dataset and called it a mini-batch. Doing this help us achive the advantages of both the former variants we

Sol. So, after creating the mini-batch of fixed size, we do the following steps in one epoch:

1) pick a mini batch
2) feed it to Neural Network
3) calculate the mean gradient of the mini-batch
4) use the mean gradient we calculated in step 3 to update the weights
5) Repeat step 1-4 for the mini-batches we created

Q.20 What is Regularization? Explain any two regularization techniques.

→ Regularization :-

Regularization refer to a set of different technique that lower the complexity of a neural network model during training, and thus prevent the overfitting

L2 & L1 regularization

L1 and L2 are the most common types of Regularization. These update the general cost function by adding another term known as the regularization term

cost function = Loss (say, binary cross entropy) + Regularization term

Due to addition of this regularization term the value of weight matrices decrease because it assumes that a neural network with smaller weight matrices leads to simpler models. Therefore, it will also reduce overfitting to quite an extent

However, this regularization term differs in L1 and L2

In L2, we have

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} \cdot \sum \|w\|^2$$

Here, lambda is the regularization parameter. It is the hyperparameter whose value is optimized for better result. L2 regularization is also known as weight decay as it forces the weight to decay towards zero (but not exactly zero)

In L1, we have

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} * \sum \|w\|$$

In this, we do penalize the absolute value of the weights. Unlike L2, the weights may be reduced to zero here. Hence it is very useful when we are trying to compress our model. otherwise, we usually prefer L2 over it.