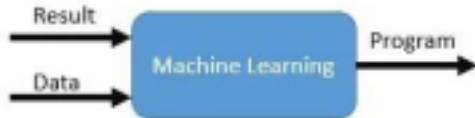**Name :- Ajit Patel**
Roll No. :- 0827CI191005
Submitted To: Dr. Shilpa Bhalerao

Set-A

Q1) a What distinguishes Machine Learning from Traditional/Conventional Programming? How we can broadly categorize various machine learning algorithms. (5 Marks)

Conventional Progamming	Machine Learning
In conventional programming, programs are created manually by providing input data and based on the programming logic, and the computer generates the output.	In machine learning programming, the input and output data are fed to the algorithm, creating the program.
Conventional Programming uses conventional procedural language. It could be assembly language or a high-level language such as C, C++, Java, JavaScript, Python, etc	Artificial intelligence is an umbrella term that contains many realms like machine learning, image processing, neural networks, cognitive science, and many more.
Conventional programming is a manual process, which means the programmer creates the logic of the program. They need to code the rules and write lines of code manually	In machine learning language, the computer uses a pre-written algorithm and learns how to solve the problem itself. It is a more sophisticated way of solving a problem.

b. Distinguish between Classification and Regression with proper illustrations? (4 Marks)

Regression Algorithm Classification Algorithm	
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.

In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

c. What is the key idea behind Simple Linear Regression? When should we use multinomial Linear Regression (4 Marks)

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know

A) How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
B) The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Simple linear regression is a parametric test, meaning that it makes certain assumptions about the data. These assumptions are:

A) Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
B) Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
C) Normality: The data follows a normal distribution.
D) The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Q 2 a.When do we say a model is overfitted or underfitted? What are some options for dealing with it?

Overfitting : It refers to a model that models the training data too well.

• Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

• A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Underfitting: It refers to a model that can neither model the training data nor generalize to new data.

- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.Underfitting is easy to detect given a good performance metric.
- Underfitting can be avoided by using more data and also reducing the features by feature selection.

b. Why sigmoid activation function suffers from Vanishing Gradient problem. What is the effect if gradient vanishes?

Vanishing Gradient Problem: As more layers using certain activation functions are added to neural networks, the gradients of the loss function approaches zero, making the network hard to train.

- The sigmoid function, squishes a large input space into a small input space between 0 and 1. Therefore, a large change in the input of the sigmoid function will cause a small change in the output. Hence, the derivative becomes small.

- A small gradient means that the weights and biases of the initial layers will not be updated effectively with each training session. Since these initial layers are often crucial to recognizing the core elements of the input data, it can lead to overall inaccuracy of the whole network.

b. How can we rescale the data using Normalization and Standardization while doing feature engineering?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

- μ is the mean of the feature values and is the standard deviation of the feature values. Note ↻

  that in this case, the values are not restricted to a particular range.

c. Determine the difference between linearity and nonlinearity. What are limitations of Perceptron Model?

Linear algorithms assume, that the sample features x and the label output y are linearly related and there is an affine function $f(x)=\langle w,x\rangle+b$ describing the underlying relationship.

Nonlinear algorithms assumes a nonlinear relationship between x and y. Thus, $f(x)$ can by a function of arbitrary complexity.

A perceptron model has limitations as follows:

- The output of a perceptron can only be a binary number (0 or 1) due to the hard limit transfer function.

- Perceptron can only be used to classify the linearly separable sets of input vectors. If input vectors are non-linear, it is not easy to classify them properly.

3.a Design a Neural Network with two hidden layers containing 5 neuron each. The dataset to be classified contains 10 features for spam/ham detection. Calculate how many parameters will be learned by model (including weights and bias) (4 Marks)
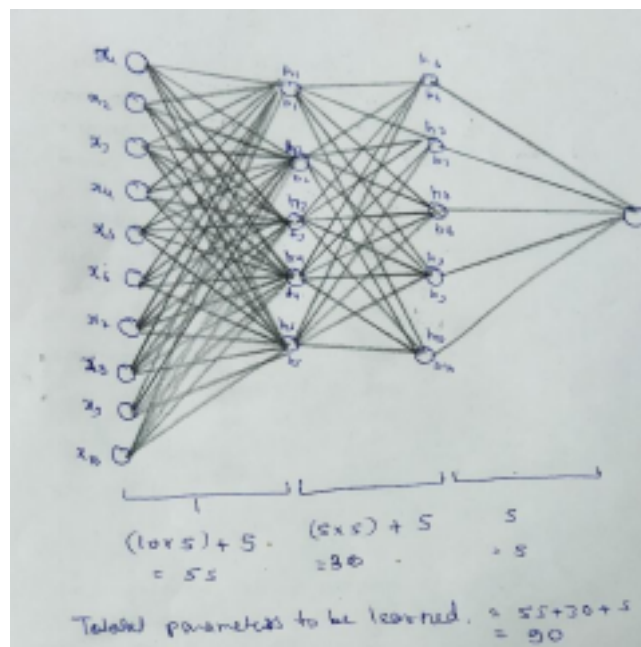
b. What is Convergence in context of Machine learning? (4 Marks)

A model is said to converge when the series $s(n)=loss_{wn}(\hat{y},y)$ (Where wn is the set of weights after the n'th iteration of back-propagation and s(n) is the n'th term of the series) is a converging series. The series is of course an infinite series only if you assume that loss = 0 is never actually achieved, and that learning rate keeps getting smaller.

Essentially meaning, a model converges when its loss actually moves towards a minima (local or global) with a decreasing trend. Its quite rare to actually come across a strictly converging model but convergence is commonly used in a similar manner as convexity

c. What is gradient descent in machine learning? Write step by step gradient descent algorithm.

(4 Marks)

Gradient Descent is known as one of the most commonly used optimization algorithms 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.

Gradient Descent method's steps are:

- choose a starting point (initialisation)
- calculate gradient at this point
- make a scaled step in the opposite direction to the gradient (objective: minimise) •
repeat points 2 and 3 until one of the criteria is met:
- maximum number of iterations reached
- step size is smaller than the tolerance.

………
… ………………

Set-B

Q1 a. What is a Neural Network and how does it replicate human brain functionality? (5 Marks)

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Neural networks are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network

b. Is cross-entropy loss function better than MSE for classification problems? Justify. (4 Marks)

- Cross-entropy (or softmax loss, but cross-entropy works better) is a better measure than MSE for classification, because the decision boundary in a classification task is large (in comparison with regression).

- Sigmoid +mse will converge slower compare sigmoid+cross entropy due to the gradient vanishing issue

2 a.Why learning rate is considered as an important hyper-parameter ? What is it's significance? (4 Marks)

The learning rate hyperparameter controls the rate or speed at which the model learns. Specifically, it controls the amount of apportioned error that the weights of the model are updated with each time they are updated
Learning rate is a scalar, a value that tells the machine how fast or how slow to arrive at some conclusion. The speed at which a model learns is important and it varies with different applications. A super-fast learning algorithm can miss a few data points or correlations which can give better insights into the data. Missing this will eventually lead to wrong classifications.

And, to find the next step or the adjacent data point, the gradient descent algorithms multiply the gradient by learning rate (also called step size).

For example, if the gradient magnitude is 1.5 and the learning rate is 0.01, then the gradient descent algorithm will pick the next point 0.015 away from the previous point.

b. The centriod of the sigmoid activation function is not zero. What is the effect of it during model learning? (4 Marks) OR What is the principle behind KNN? What are its application area? (4 Marks)

c. What is 'training Set' and 'test Set' in a Machine Learning Model? How Much Data will you allocate for your Training,Validation, and Test Sets? (4 Marks)

Training set : It is the set of data that is used to train and make the model learn the hidden features/patterns in the data. In each epoch, the same training data is fed to the neural network repeatedly, and the model continues to learn the features of the data.

The Test Set: The test set is a separate set of data used to test the model after completing the training.It provides an unbiased final model performance metric in terms of accuracy, precision, etc.

Allocating training, validating and Test sets:

There are two major concerns while deciding on the optimum split:

- If there is less training data, the machine learning model will show high variance in training.
- With less testing data/validation data, your model evaluation/model performance statistic will have greater variance.

Essentially, you need to come up with an optimum split that suits the need of the dataset/model.

3 a. Design a Neural Network with two hidden layers containing 3 neurons each. The dataset to be classified contains 10 features for handwritten digit classification .Calculate how many Parameters will be learned by model (including Weights and bias) (4 Marks)
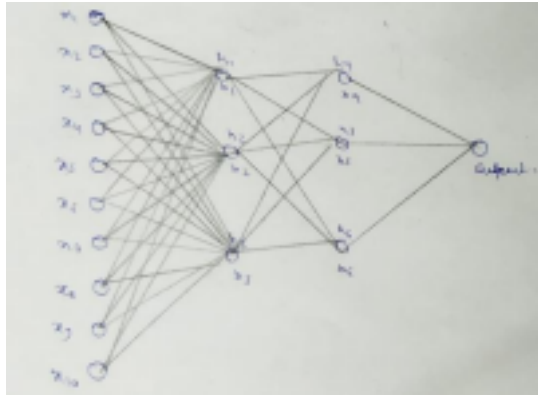
b Explain Correlation and its implication in Machine Learning.(4 Marks)

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

Positive Correlation: Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.
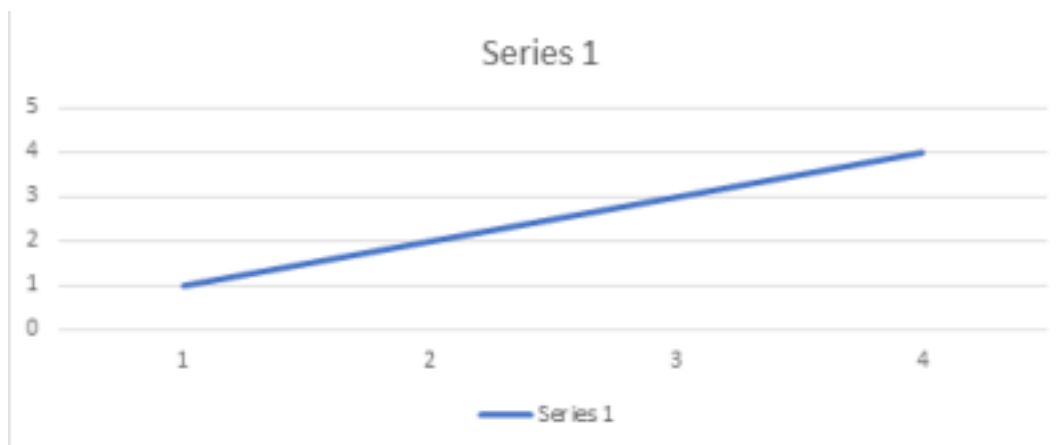
Negative Correlation: Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.
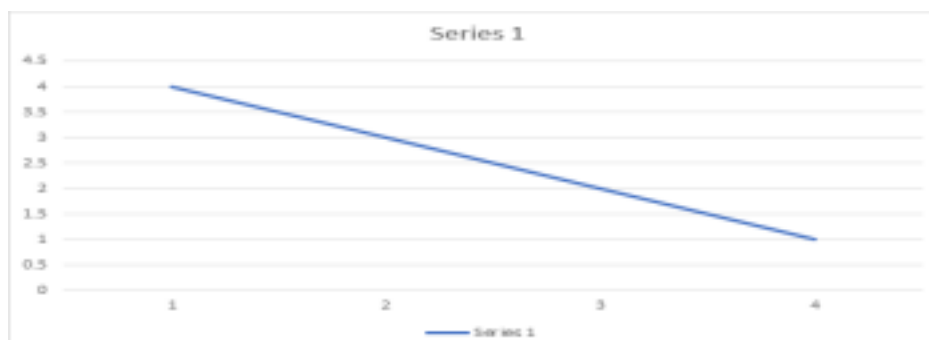
No Correlation: Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable(s) doesn't increase or decreases.

c. Explain the role of activation function in Neural Network. Give examples of various activation functions. Explain any one. (4Marks)

The activation function is the most important factor in a neural network which decided whether or not a neuron will be activated or not and transferred to the next layer. This simply means that it will decide whether the neuron's input to the network is relevant or not in the process of prediction. For this reason, it is also referred to as threshold or transformation for the neurons which can converge the network.

Activation functions help in normalizing the output between 0 to 1 or -1 to 1. It helps in the process of backpropagation due to their differentiable property. During backpropagation, loss function gets updated, and activation function helps the gradient descent curves to achieve their local minima.

Different activation functions:

  • Linear: Linear is the most basic activation function, which implies proportional to the input. Equation Y = az, which is similar to the equation of a straight line. Gives a range of activations from -inf to +inf. This type of function is best suited to for simple regression problems, maybe housing price prediction
  • ReLU
  • ELU
  • PReLU
  • Sigmoid
  • Softmax
  • Tanh
  • Swish