Leading University

Department of Computer Science and Engineering

Course Title: Machine Learning

Course Code: CSE - 4233

Submitted to:

Shafkat Kibria
Assistant Professor

Department of CSE

Submitted by:

Anisa Jahan

ID : 1912020171

Dept: CSE

Batch: 50th

Section: 11(D)

Date of Submission : 19/10/2022

Ans No: 1

MIPS: Million instructions per second (MIPS) is an approximate measure of a computer's raw processing power. MIPS figures can be misleading because measurement techniques often differ and different computers may required different sets of instructions of perform the same activity.

Ans No: 2

There are 10^{11} neurons and 10^{12-13} synapses in the brain.

Ans No: 3

Moravec's paradox: Moravec's paradox is a phenomenon that occurs when we use AI-powered tools. It observes that humans find complex tasks, simple for AI to learn. That is when compared to simple sensorimotor skills that human possess naturally AI, for example can solve complex logical issues and perform advanced mathematics.

Ans No: 4

Machine Learning: It is a branch of Artificial Intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Another definition,

"Field of study that gives the computers the ability to learn without being explicitly programmed"

—Arthur Samuel (1959)

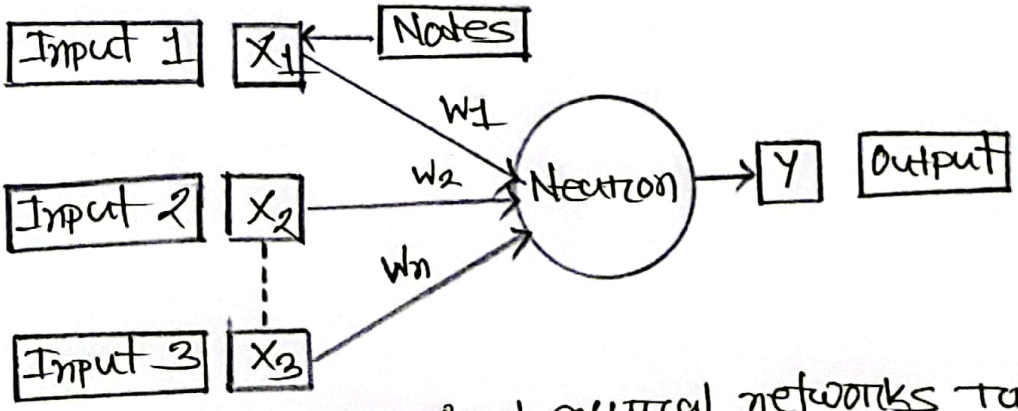
Machine learning is important because it gives enterprises a view of trends in customer behaviour and business operation patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations.

Ans No: 5

Some of ML methods :

⊞ Artificial neural networks (ANN) : The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes,

The typical Artificial neural network looks something like the given figure :



Dendrites from Biological neural networks represents inputs in Artificial neural networks, cell nucleus represents nodes, synapse represents weights and

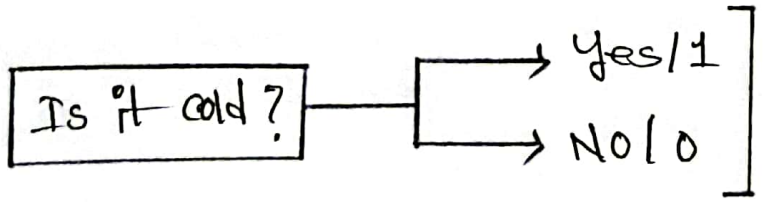
Axon represents output.

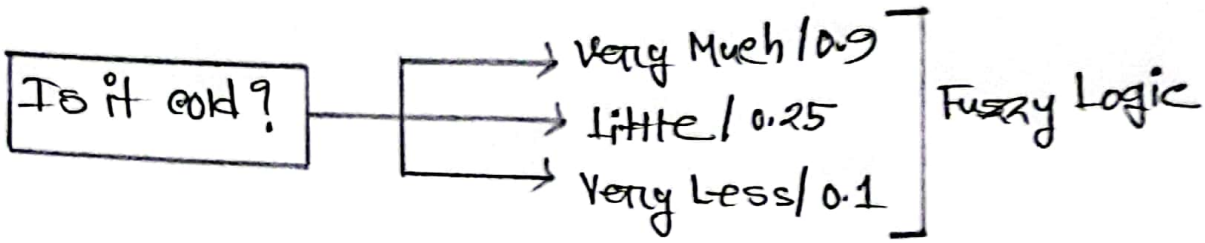
Artificial neural network primarily consists of three layers:

1. Input Layer
2. Hidden Layer
3. Output layer.

Fuzzy logic: It is a method of reasoning that resembles human reasoning. The approach of FL imitates the way to decision making in humans that involves all intermediate possibilities between digital values YES and NO.

The conventional logic block that a computer can understand takes precise input and produce a definite output as TRUE or FALSE, which is equivalent to human's YES or NO.





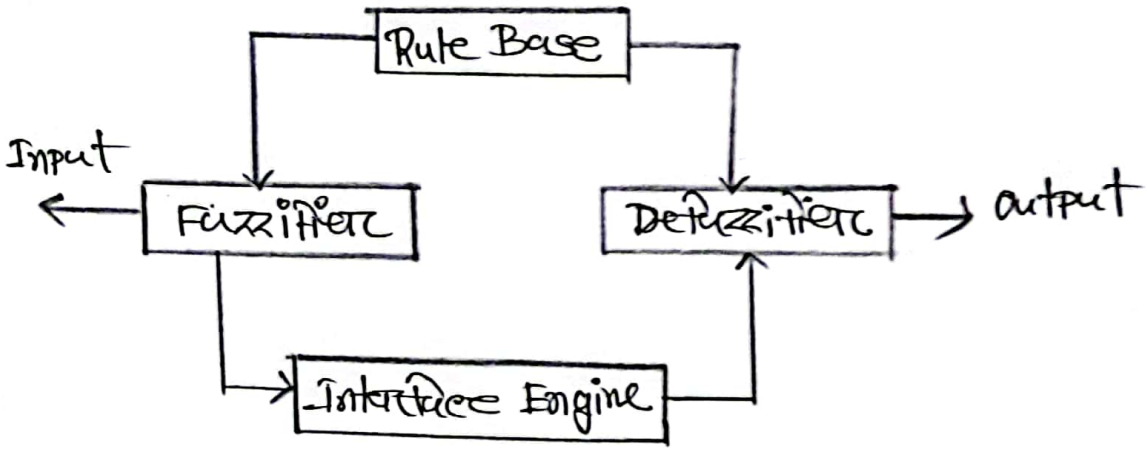


Fig: Fuzzy Logic Architecture.

Genetic Algorithm: Genetic Algorithms are adaptive heuristic search algorithms that belong to the larger part of evolutionary algorithm. Genetic algorithms are based on the ideas of natural selection and genetics. They are intelligent exploitation of random search provided with historical data to direct the search into the region of better performance in solution space. They are commonly used to generate high-quality solutions for optimization problems and search problems.

Support Vector Machine (SVM): It is one of the most popular supervised Learning algorithms, which is used for classification as well as Regression problems. However, primarily, it is used for classification problems in Machine Learning.

The goal of SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Ans No: 6

Some example of ML based application:

1. Traffic Alerts
2. Image Recognition
3. Video surveillance
4. Sentiment Analysis
5. Product Recommendation.

Ans No: 7

Supervice learning is the types of machine in which machines are trained using we "labelled" training data, add on ~~based~~ basis of that data, machines predict the output. The labelled data means some input data is already tagged with 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.

For example, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

- If the shape of the object is round and has a depression at the top, is red in color, then it will be labeled as: Apple

- If the shape of the object is a long curving cylinder having green yellow color, then it will be labeled as: Banana.

Now suppose after training the data, you have given a new separate fruit, say Banana from basket and asked to identify it, since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data and then applies the knowledge to test data.

Ans No : 8

It is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

Here the task of a machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

The goal of unsupervised learning is find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

For example, suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never it does not have any idea about the features of the dataset. The task of the unsupervised algorithm is to identify images features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images. It mainly deals with unlabelled data.

Ans No: 9

A priori probability refers to the likelihood of an event occuring when there is a finite amount of outcomes and each is equally likely to occur.

The outcomes in a priori probability are not influenced by the prior outcome.
In the other words, a priori probability is derived from logically examining a event.

A priore probebility $= \frac{f}{N}$

where,

- f refers to the no of desirable outcomes
- N refers to the total no. of output.

Ans No: 10

A posterior probability, in Bayesian statistics, is the revised or updated probability of an event occurring after taking into consideration new information. The poster probability is calculated by updating the prior probability using Bayes' theorem.

Ans No: 11

The hypothesis is a common term in Machine Learning and data science projects. As we know, machine learning is one of the most powerful technologies across the world, which helps us to

Predict results based on past experience. Moreover, data scientists and ML professionals conduct experiments that aim to solve a problem. These ML professionals and data scientists make an initial assumption for the solution of the problem.

Ans No: 12

Univariate data is the type of data in which the result depends only on one variable. For instance, dataset of points on a line can be considered as a univariate data where abscissa can be considered as input feature and ordinate can be considered as output.

Univariate linear regression focuses on determining relationship between one independent variable and one dependent variable. There is only one input feature vector. The line of regression will be in the form of:

$$y = b_0 + b_1 x$$

where, b_0 and b_1 are the coefficients of regression.

since, it is being tried to predict regression coefficient b_0 and b_1 by training a model.

Ans No : 13

A cost function is an important parameter that determines how well a machine learning model performs for a given dataset. It calculates the difference between the expected value and predicted value and represents it as a single real number.

Ans No : 14

Gradient Descent is an iterative optimization algorithm used in Machine Learning to minimize cost function.

Ans No : 15

Gradient Descent is first-order iterative optimization algorithm for finding local minimum of a differentiable function.

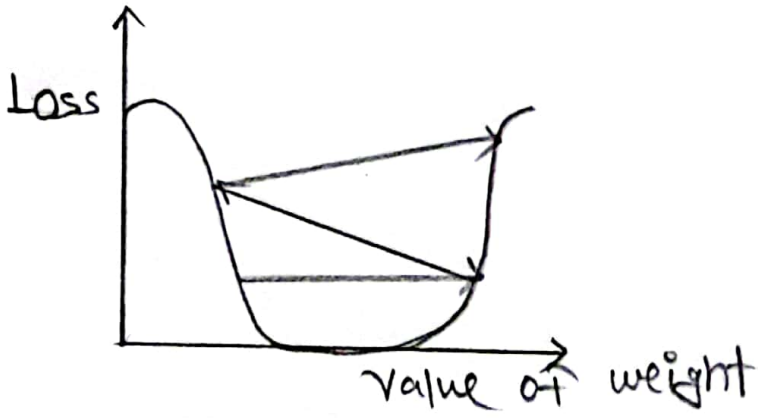
Gradient Descent,

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Where,

α = Learning rate.

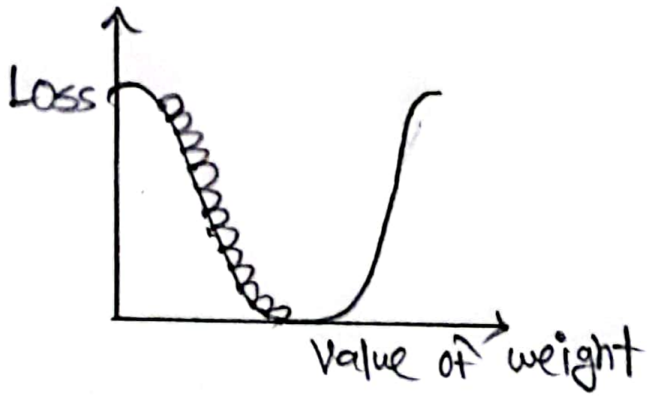
- If the learning rate is high, it results in larger steps but also leads to risks of overshooting the minimum. That means, it may not reach the local minimum because it bounces back and forth between the convex function of gradient descent. It may fail to converge or even diverge.



For example, learning rate (α) is high like 0.5 or 0.7. In this case the major portion of gradient of current sample will be used for weight updation. Now for the next sample again the new gradient will be generated which is different from the previous one and the major portion (as the learning rate is high as 0.5)

is going to be used for weight (o_j) updation. There are good chances that the weight updation caused by previous gradient may adversely get effected by this current gradient

- If we set the learning rate (α) to a very small value, gradient descent will eventually reach the local minimum but that may take a while.



For example, If the gradient of learning rate is very is very small as 0.0001 or 0.00001 then very small portion of the gradient will be used for weight updation. The same will be true for the next samples. And at the end the networks learns very slowly. It requires many number of epochs to learn the pattern

Ans No : 16

If the gradient descent converge to local or global minimum, then do not update θ value.

For convex problems, gradient descent can find the global minimum easily. The slop of the cost function is at zero or just close at zero, the model stops learning further. Local minimum generates the shape similar to the global minimum, where the slop of the cost function increases on both sides of the current points.

Ans No : 17

We should choose learning rate α "fixed" because Gradient Descent can converge to a local minimum, even with the learning rate α fixed.

$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

As we approach a local minimum gradient descent will automatically take smaller steps. So, no need to decrease α over time.

Ans No: 18

Univariate	Multivariate
1. Univariate linear regression focusses on determining relationship between on independent variable.	1. Multivariate linear regression has more at least two independent variable.
2. The information deals with only one quantity that change.	2. The information deals with multi quantity those change.

Ans No: 19

Feature scaling is a method used to normalize the range of independent variable or features of data. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

we need feature scaling, because —

- scaling the features makes the flow of gradient quickly reach the minima of the cost.
- without scaling features, the algorithm may

be baised toward the feature which has value higher in magnitude. Hence we scale features that bring every feature in the same range, and the model uses every feature wisely.

Ans No: 20

No, we should not feature scaling in one variable in one linear regression.

For example, to find the best parameter for a linear regression model, there is a closed-form solution, called the Normal Equation. If our implementation makes use of the equation there is no stepwise optimization process. So, feature scaling is not needed.

Ans No: 24

After applying feature scaling, the approximate range for every feature is $-1 \leq x_i \leq 1$.

Ans No : 22

We know,

Mean normalization, $x_i = \frac{x_i - \mu_i}{s_i}$

where, x_i = Feature

μ_i = Average value of ~~feature~~

s = standard deviation

$= \max(x_i) - \min(x_i)$

Given,

Age of house $= x_i$

$\mu_i = 38$

$\max(x_i) = 50$

$\min(x_i) = 30$

$\therefore x_i' = \frac{x_i - \mu_i}{s_i}$

$= \frac{\text{age of house} - 38}{\max(x_i) - \min(x_i)}$

$= \frac{\text{age of house} - 38}{50 - 30}$

$= \frac{\text{age of house} - 38}{20}$

Ans No: 23

The gradient descent algorithms purpose is to minimize a given function (cost functions) A good way to make sure the gradient descent algorithm works properly is by ploting the cost function as the optimization runs. putting the number of iterations on the x-axis and the value of the cost function on the y-axis. This help us see the value of our cost function after each iteration of gradient descent and provides a way to easily spot how appropriate ~~a way to easily~~ our learning rate is. The left figure below shows such a plot, while the figure on the right illustrates the difference between good and bad learning rate.

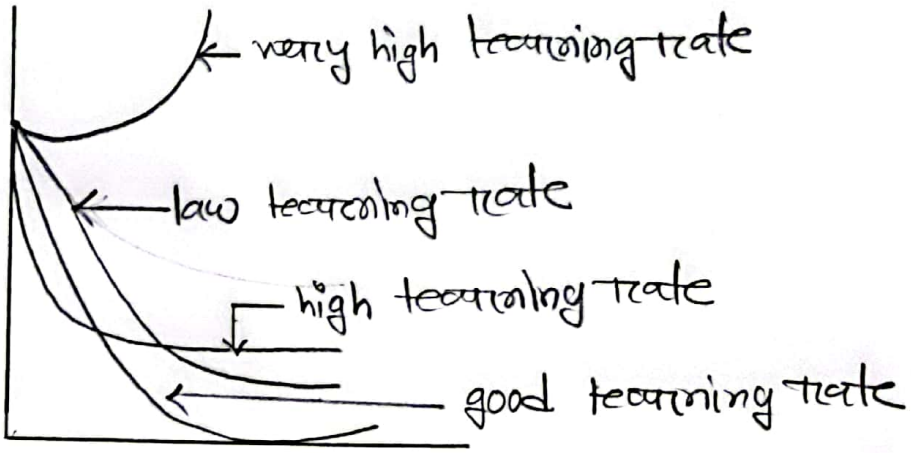
so, we can say that if the gradient descent algorithm is working properly, the cost function should decrease after every iteration.
When gradient descent can't decrease the cost function anymore and remain more or less on the same level, it has converged.

Ans No: 24

We know if gradient descent works properly the cost function should decrease after each iteration. But when gradient descent can no longer reduce the cost function and remains at more or less the same level, it has converged and it may sometimes require a large change in the number of iterations to converge.
If the plot shows that the learning curve just goes up and down, without really reaching a lower point. Then we try to reduce the learning rate. Also try 0.001, 0.003, 0.01

0.03, 0.1, 0.3, 1 etc as the learning rate and see which one performs best.

As shown in the figure below:



Ans No: 25

We will automatically choose the learning rate α to be (0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1...) and plot the $J(\theta)$ against the number of iterations for each to see which value of α gives the best convergence.

Ans No: 26

Feature engineering is the pre-processing step of machine of machine learning, which is used to transform raw data into feature that can be used for creating a predictive model using Machine learning or statistical modeling. The aim of feature ~~learning~~ engineering is to prepare an input data set that best fits the machine learning algorithm as well as to enhance the performance of ML models. We need feature engineering because:

- Better features mean flexibility
- Better features mean simpler models
- Better features mean better results

Ans No: 27

Polynomial regression is a regression algorithm that models the relationship between a independent(x) and dependent variable (y) as nth degress polynomial. The polynomial regression eqn below:

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + \dots + b_n x_1^n.$$

We need polynomial regression, because -

- If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in simple linear regression but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase the error rate will high, and accuracy will be decreased so, for such cases, where data points are arranged in a non-linear fashion, we need the polynomial regression model.

Ans No: 28

Logistic Regression is an example of supervised learning that used to calculate or predict the probability of a binary event occurring.

sometimes our data contains outlines, in which case our linear regression fails to produce a line of best fit based on the data points.

so in regression, we are producing continous values but what if we want to predict categorically values like true or false, correct or incorrect, yes or no then our linear regression model does not work, so the solving this type of problem will required us to use logistic regression.

Ans No: 29

sigmoid function is used for the logistic regression

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \quad ; \quad 0 \le h_\theta(x) \le 1$$

or, $$g(z) = \frac{1}{1+e^{-z}}$$

Ans No: 30

A decision boundary, is a surface that separates data points belonging to different class lables,

Ans No: 31

Hypothesis function for non-linear decision boundaries,

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

Ans No: 32

If we try to use the cost function of the linear regression in "Logistic Regression" then it will be of no use as it will end up as non-convex function with many local minima, where it will be very difficult to minimize the cost value and find the global minimum.

So, logistic regression, the cost function is

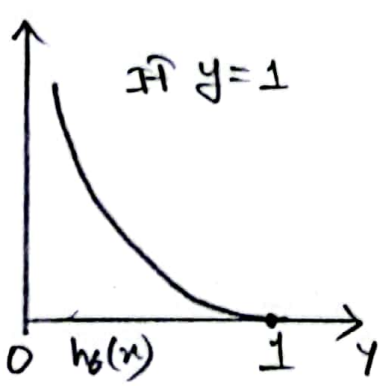
defined as,

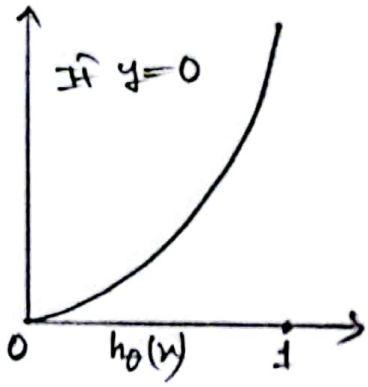
$$\text{cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & ; \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & ; \text{if } y = 0 \end{cases}$$

Ans No: 33

The cost function for logistic regression:

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





Ans No: 34

Let,

$R = \{\text{Red Balls}\}$

$B = \{\text{Blue Balls}\}$

$G = \{\text{Green Balls}\}$

$$P(R) = \frac{50}{125} = 0.3846$$

Similarly, $P(B) = 0.3461$, $P(G) = 0.2692$

And,

$P(T/R) = 0.4$, $P(T/B) = 0.266$, $P(T/G) = 0.333$

Now,

$$P(R/T) = \frac{P(T/R)\ P(R)}{P(T/R)\ P(R) + P(T/B)\ P(B) + P(T/G)\ P(G)}$$

$$= \frac{0.4 \times 0.3846}{(0.4 \times 0.3846) + (0.266 \times 0.3461) + (0.333 \times 0.2769)}$$

$$= 0.46$$

$$P(B/t) = \frac{0.2666 \times 0.3461}{(0.4 \times 0.3846) + (0.2 \times 0.3461) + (0.333 \times 0.269)}$$

$$= 0.295$$

$$P(G/t) = \frac{0.333 \times 0.269}{(0.4 \times 3846) + (0.2 \times 0.3461) + (0.333 \times 0.269)}$$

$$= 0.267.$$