We at The Data Monk hold the vision to make sure everyone in the IT industry has an equal stand to work in an open domain such as analytics. Analytics is one domain where there is no formal under-graduation degree and which is achievable to anyone and everyone in the World.

We are a team of 30+ mentors who have worked in various product-based companies in India and abroad, and we have come up with this idea to provide study materials directed to help you crack any analytics interview.

Every one of us has been interviewing for at least the last 6 to 8 years for different positions like Data Scientist, Data Analysts, Business Analysts, Product Analysts, Data Engineers, and other senior roles. We understand the gap between having good knowledge and converting an interview to a top product-based company.

Rest assured that if you follow our different mediums like our blog cum questions-answer portal www.TheDataMonk.com , our youtube channel - The Data Monk, and our e-books, then you will have a very strong candidature in whichever interview you participate in.

There are many blogs that provide free study materials or questions on different analytical tools and technologies, but we concentrate mostly on the questions which are asked in an interview. We have a set of 100+ books which are available both on Amazon and on The Data Monk e-shop page

We would recommend you to explore our website, youtube channel, and e-books to understand the type of questions covered in our articles. We went for the question-answer approach both on our website as well as our e-books just because we feel that the best way to go from beginner to advance level is by practicing a lot of questions on the topic.

We have launched a series of 50 e-books on our website on all the popular as well as niche topics. Our range of material ranges from SQL, Python, and Machine Learning algorithms to ANN, CNN, PCA, etc.

We are constantly working on our product and will keep on updating it. It is very necessary to go through all the questions present in this book.

Give a rating to the book on Amazon, do provide your feedback and if you want to help us grow then please subscribe to our Youtube channel.

ARTIFICIAL NEURAL NETWORKS

Q1. What is the idea behind neural network?

A1. The idea of neural network in machine learning is taken from our brain cell. The basic idea behind it is to copy our brain cell .To act like our brain cells and solve problems.

Q2. What is the basic concept of Neural Network?

A2. In neural network the basic process is forward propagation than calculating the loss function and backward propagation.

Q3. What is forward propagation?

A3. Suppose we have inputs as x1,x2,x3 and weights as w1,w2,w3 . Firstly, the inputs will be passed to the hidden layer with their respective weights and bias.

y=x1w1+x2w2+x3w3 + bias ,

In hidden layer, activation function is applied to the input(y).

Z=act(y),

There are different kinds of activation function.

Now when Z is passed to the output layer, the predicted output is compared to the actual output.

Q4. What is activation function?

A4. Activation functions in neural networks are used to make the input (y) from a neuron is transformed in the output from a node/nodes ranging according to the type of activation function applied. For ex. If we apply sigmoid the range is 0 to 1.

There are different types of activation function. There are three layers in the neural network: input layer, hidden layer, output layer. Mostly, same activation function is applied to every hidden layer. But in output layer a different activation function is applied and it is selected on the basis of what type of prediction is required by the model. For example, if someone puts a hot object on your hand. So that neurons will get activated and send the signals to the brain to react on that.

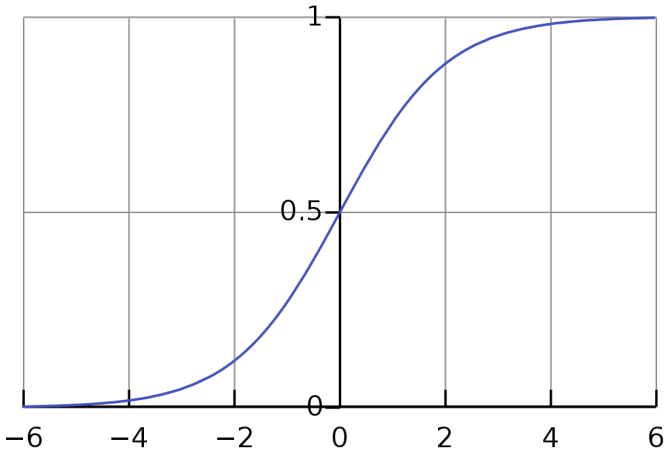
Q5. Different types of activation functions?

A5. There are many types of activation functions:

- Sigmoid activation function
- ReLU activation function
- Leaky ReLU activation function
- PReLU activation function
- ELU activation function
- Softmax activation function
- Tanh activation function

Q6. What is a Sigmoid activation function?

A6. You have studied about this activation function earlier in logistic regression algorithm. Sigmoid is basically an “S” liked curve, in which we set a threshold value. So if the input value is greater than threshold value than only the activation function will get activated otherwise it will get deactivated.



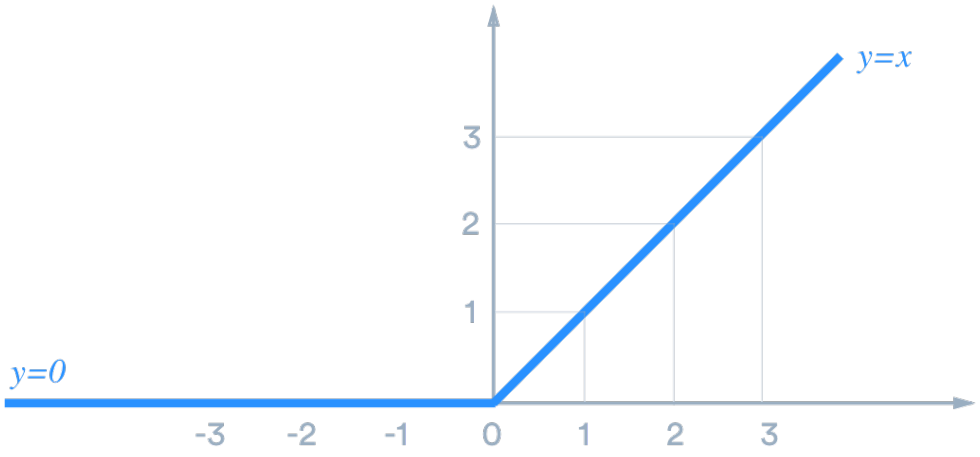
It is represented by the formula:

Sigmoid = 1 / (1 + e^-z)

Sigmoid activation function is basically used in the output layer.

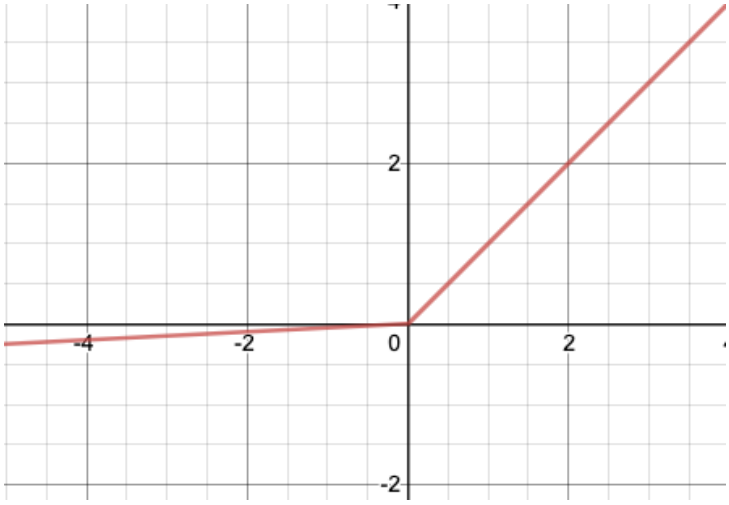
Q7. What is a ReLU activation function?

A7. ReLU stands for Rectified Linear Unit. ReLU activation function works on the formula of “max(z,0)”.We take the max number out of the two. If the input value is smaller than 0 than output will be “0”. If the input value is greater than “0” than the output will be considered as that value.



Q8.What is Leaky ReLU activation function?

A8. There is a limitation in the ReLU activation function. The limitation is that when we calculate the derivative of the ReLU activation function than the derivative of values less than “0” is “0”.Now suppose we have a large weight and we are calculating the derivative to adjust the weights through chain rule and one of the derivative comes out to be “0” in the chain rule. Than the whole derivative will become “0” and it will considered as a dead neuron. So to overcome this problem we use Leaky ReLU activation function. Leaky ReLu adds a small value to the derivative for value less than “0”, the value will be small like “0.01”. Due to which it will not let the derivative value be “0”.



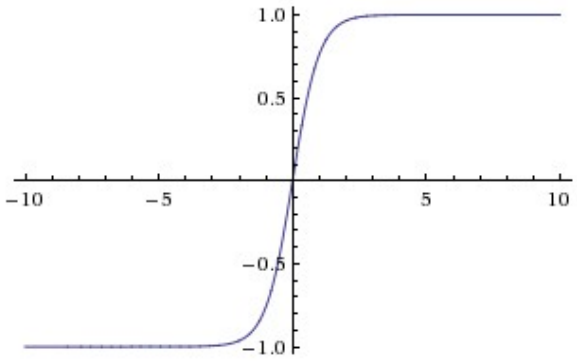
Q9. What is softmax activation function?

A9. Softmax activation function is mostly used in the time when there is multiple classification problem. It is represented by the formula given below. It is mostly used in the output layer.

$$\phi(z) = \frac{e^i}{\sum_{j=0}^{k} e^j} \quad \text{where } i=0,1,\ldots.k$$

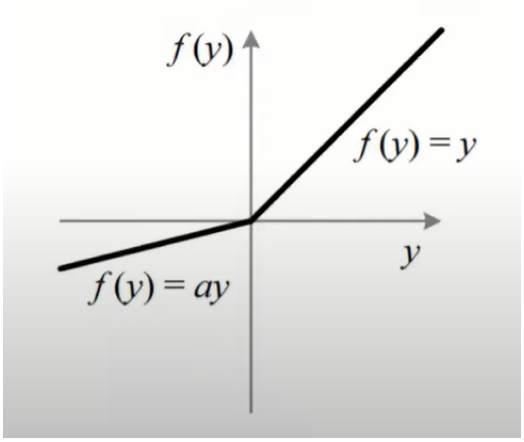
Q10. What is tanh activation function?

A10. Tanh activation function is similar to the sigmoid activation function. The only difference is that tanh is a zero-centric activation function. Due to which it is better than sigmoid activation function.



Q11. What is PReLu activation function?

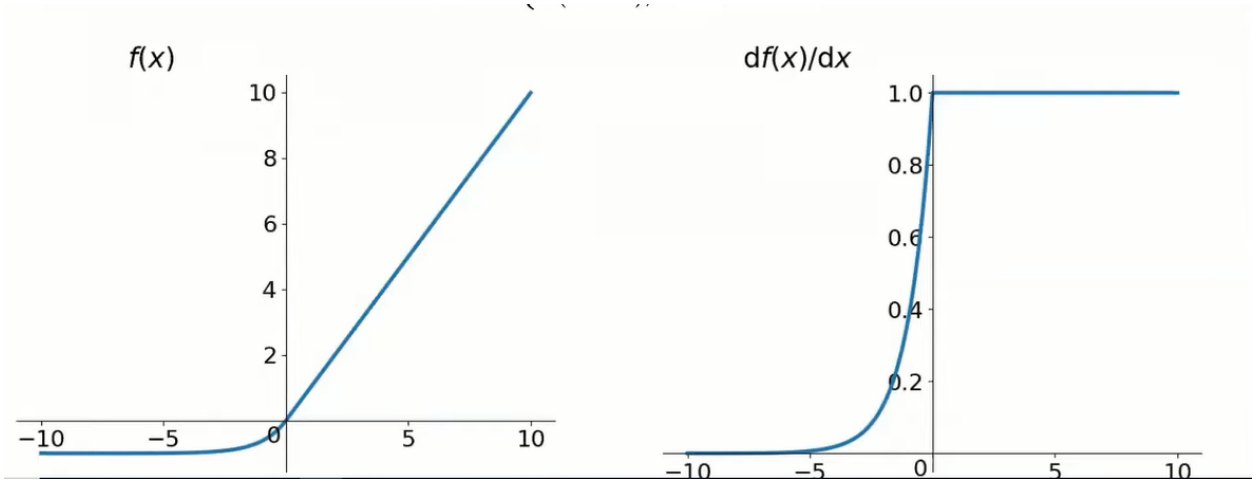
A11. PReLU is basically parametric ReLU activation function. It is kind of an improved version of ReLU. Basically, it has a small slope in the negative region which prevents the problem of dead neuron.



$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases}$$

Q12. What is ELU activation function?

A12. ELU stands for Exponential Linear Unit. It is an updated kind of ReLU . In this we add a hyper parameter to the values less than “0”, whose derivative comes out to be “0”. But after applying hyper parameter to it than the derivative will not be “0”. But it has a disadvantage that it is computationally expensive.



$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{otherwise} \end{cases}$$

Q13. What would happen without activation function?

A13. If we do not apply any activation function to the inputs then the value of (y) will be a large value and it will be becoming a larger value after every layer. It will not be easy to calculate that values, it will be requiring very high computational power. So, we use an activation function to sum down the values.

Q14. What is loss function?

A14. When we reach the output layer with a predicted output, now we need to compare it with the actual output. So here we use the loss function we calculate the difference between the actual output and predicted value. If the difference is large then we will adjust the weights and try to reduce it as much as possible.

Q15. What are different types of loss functions?

A15. There are different types of loss functions to handle different types of variables

- Regression
- Single class classification
- Multi class classification

Q16. What is the loss function used in regression problem?

A16. Different types of loss function in regression problem are:

- Mean Square Error Loss :

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Where yi is actual output and yi^ is predicted output.

Advantages:

1. The Mean Squared Error penalizes the model for making large errors by squaring them.
2. When we solve the above equation we will get a quadratic equation:
   (i). When we plot the graph for quadratic equation , we will get a Gradient Descent with only 1 global minima.
   (ii). We do not get any local minima

Disadvantages:

1. It is not robust to the outliers. Meaning, whenever we have an outlier it will not function properly and leave errors.

- Mean Absolute Error Loss :

$$\underbrace{\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} \underbrace{|y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i|}_{\text{actual value}}$$

When we solve the above equation, we will get a linear equation.

Advantages:

1. It is more robust to outliers as compared to the mean square error loss.

- Huber Loss :

$$Hubber\ Loss = \begin{cases} \frac{1}{2}(y - y_p)^2, & |y - y_p| \le \delta \\ \delta|y - y_p| - \frac{1}{2}\delta^2, & |y - y_p| > \delta \end{cases},$$

where , y is actual output , yp is predicted output and dell symbol is the hyper parameter.

Huber Loss is basically a combination of mean squared error and mean absolute error. As we know if we have outliers than the quadratic equation will not work nicely. So we splitted the cases in Huber loss , if we have outliers in the dataset than we will use linear equation i.e. mean absolute error loss and if we do not have outliers in the dataset than we will use quadratic equation i.e. mean square error loss.

Q17. What is the loss function used in classification problem?

A17. The widely used loss function in the classification problem is Cross Entropy. It is basically makes use of sigmoid function as we studied in logistic regression.

It is represented by the formula:

loss = -y * log(yp) – (1-y) * log(1 – yp)

Where, y is actual output and yp is predicted output.

We calculate the yp by using sigmoid function:

yp = sigmoid = 1/ (1 + e^(-z))

Q18.What is the loss function used in multi-class classification problem?

A18. In multi class classification problem we use Multi Class Cross Entropy Loss .

It is represented by the formula:

$$L(\hat{y}, y) = -\sum_{k}^{K} y^{(k)} \log \hat{y}^{(k)}$$

Where , y^ is predicted output , y is actual output.

So, when we have multiple classes in dependant feature than we will apply one hot encoding to it. So, the value of “k” in the above equation is 0 or 1. “1” if the element is present in that class and “0” if the element is not present in that class.

Q19. How do we classify which kind of classification problem is it?

A19. If the output features are having only two categories then it will be considered as single variable classification problem, because after the implementation of the code we will apply label encoder to the output variable so it will be converted into 0 or 1. Now if we have more than two categories than it is considered as multi-class classification problem. This happens because when we apply label encoder to the output feature then we will get the different values of the different categories.

Q20. What is cost function?

A20. It is synonymous to the loss function. The only difference is that it cost function is basically the average loss for the whole dataset.

Q21. What is Optimizer?

A21. When we calculate the loss, then we need to calculate by how much we are going to adjust the weights or how the weights must be adjusted so that our loss is minimum. Optimizer are algorithm or model used to change the attributes of the neural network such as weights and learning rate in such a way that the loss in minimized as much as possible.

Q22. What are different types of optimizers?

Q22. There are various types of optimizers:

- Gradient Descent
- Stochastic Gradient Descent(sgd)
- Mini-Batch stochastic Gradient Descent
- Stochastic Gradient Descent with momentum
- Adagrad optimizer
- Adadelta and RMS prop optimizer
- Adam optimizer

Q23.What is Gradient Descent optimizer?

A23. It is used to update the weights in the back propagation of neural network. In gradient descent we take all the inputs at a single time to perform the propagations.

It is represented by the formula:

$$Wn = Wo – n * (dL / dW)$$

Where , Wn is new/updated weight , Wo is old weight , n is learning rate.

The value of dL/dW is negative if the slope is negative or downward going and positive if the slope is positive or upward going .

Q24. What is Stochastic Gradient Descent?

A24. It is an upgraded / improved version of gradient descent. In gradient descent, suppose we have a dataset of 1000 data points we will take all the 1000 data points at a single time to compute the derivative which will take very much computational power. But in case of stochastic gradient descent, we take a single data point out of 1000 at the time of propagation.

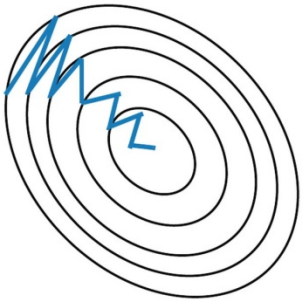
Q25. What is Mini Batch Stochastic Gradient Descent?

A25. It is and upgraded/ improved version of stochastic gradient descent. In stochastic gradient descent, we take only one data point in propagation which will use many resources to get computed and will be slow in computation. So to overcome this issue we introduced Mini Batch Stochastic Gradient Descent, in this we take “k” number of data points in the propagation which will use less resources and computational power.

Q26. What is Stochastic Gradient Descent with Momentum?

A26. It is an updated/improved version of mini batch stochastic gradient descent. When we use mini batch stochastic gradient descent we will have some noisy data. So in order to remove the noisy data we will use momentum. We will assign momentum Y=0.5(exponentially moving average). When we calculate the derivatives with the chain rule , suppose we are calculating for the W3 so with the help of momentum we will give it more importance , less importance to W2 and least importance to W2, due to which we will be able to remove some noise from the dataset.

Stochastic Gradient Descent **withhout** Momentum



Stochastic Gradient Descent **with** Momentum

Q27. What is Adagrad optimizer?

A27. Till now, we have seen that learning rate is fixed in gradient descent , stochastic gradient descent , mini batch stochastic gradient descent. So, idea behind adagrad optimizer is that it changes learning rate for every layer and neuron for every iteration. It has only one problem i.e. when number of iterations are very large than the coefficients's value will become very large . So , learning rate will become very small and it will lead to slow convergence .

Q28. What is Adadelta and RMS prop optimizer?

A28. As we know about the disadvantage of the adagrad optimizer we use Adadelta optimizer.

$Wt=W(t-1) - Nt\ (dL/dW(t-1))$

$Nt=N/\ sqrt(Wavg +E)$

$Wavg = Y^* Wavg(t-1) + (1-Y)\ (dL/dW(t))^2$

Now , also the value of Wavg will become large , but it has Y to restrict it . So, the value of Nt will be decreasing very slowly.

Q29. What is Adam optimizer?

A29. It is the most widely used and the best till date optimizer as it has the property of both the stochastic gradient descent with momentum and changing learning rate efficiently as in RMS prop.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

The first equation is with the effect of momentum and the second equation is for the change in learning rate efficiently.

So , our final formula for the adam optimizer is given below:

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Q30. What is backward propagation?

A30. After the forward propagation, we received a predicted output. Now we will compare the predicted output with the actual output .If the difference between the predicted output and actual output is a large value than we will try to minimize. To minimize the difference we use optimizers, the weights are now adjusted in such a manner that loss in minimum.

Q31. How are weights adjusted in backward propagation?

A31. The weights are adjusted with the help of optimizer function. We compare the predicted output with the actual output. We get a difference and according to this difference we adjust the weights such that this difference becomes minimum.

Q32.What is batch size?

A31. Batch size basically means that how much input values are taken in a single propagation. For example, Before lockdown we go to any place in a bulk but after lockdown we stand in a queue and we go in small batches .

Q33. What is the library used to perform neural networks?

A33. Libraries used to perform neural network:

- Tensorflow
- Sklearn
- keras

Q34.What is epochs?

A34. When we are training the neural network for one repetition with the training data. One propagation means forward propagation + backward propagation

Q35. How many layers are there in neural networks?

A35. There are basically three types of layers:

-Input layer

-Hidden layer

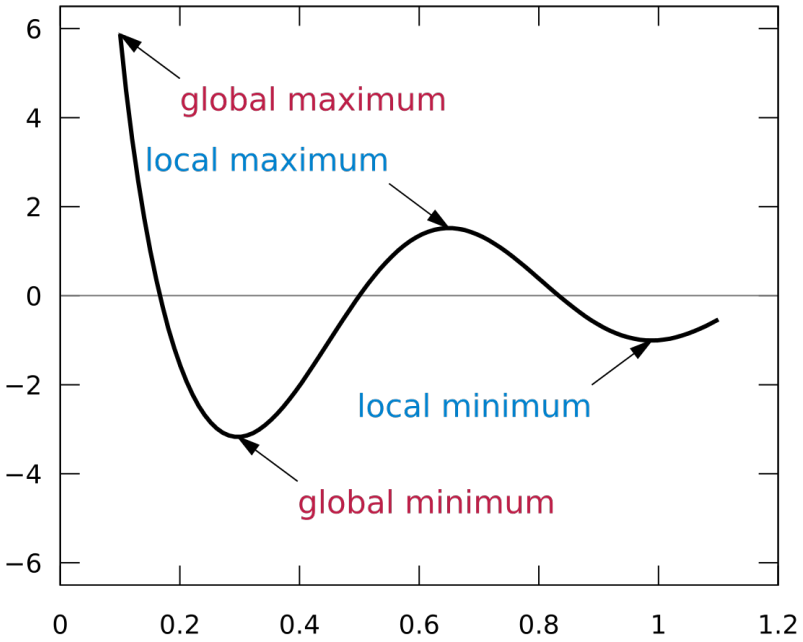
-Output layer

Q36. Which is the best optimizer?

A36. The best and most efficient optimizer till date is “adam” optimizer .As it has got the smoothness from the momentum of stochastic gradient descent and efficient changing of learning rate from the RMS prop.

Q37. What are local minima and maxima?

A37. When we reach global minima than the slope becomes “0” and it is the lowest point in the curve, but in the case of local minima, we reach a point where slope becomes “0” but it is not the lowest point in the curve. If we move forward we will reach a point which will be the lowest point in the curve (global minima).



When we reach global maxima our slope becomes “0” and it is the highest point in the curve. But in the case of local maxima we reach appoint where slope is “0” but it is not the highest point in the curve. If we move forward in the curve we will reach a higher point tan it (global maxima).

Q38. What will happen if the learning rate is set too high?

A38. If we set the learning rate of the neural network too high then it will never reach the global minima point. Because at the weight adjustment time if the learning rate is set too high then it will make the new adjusted weights a high value. (You can think of it as a hyper active scenario).

Q39. What will happen if the learning rate is set too low?

A39. If we set the learning rate of the neural network too small then it will take much more time than the usual time to reach the global minima point. Because if we set the learning rate too small then the new adjusted will the a bit different from the old value and it will take much more time to it to reach the global minima.

Q40. When does over fitting and under fitting happens in neural network?

A40. When we have a deep neural network i.e. neural network with many layers than the problem of over fitting arises. Because neural networks repeats the propagations until it gets the same predicted output as the actual output.

When we have a neural network with only one hidden layer than only the under fitting can happen.

Q41. How are weights initialized in neural network?

A41. When we initialize weights, weights should not be initialized very small , weights should not be same for every neuron , weights must haver a good variance. There are different types of weight initialization technique:

- Uniform distribution -> weight = Uniform [-1/ sqrt(ni) , 1/ sqrt(ni)]
- Xavier/ Gorat Normal -> weight = normal (0 , sqrt(2/ ni +no)]
- Xavier Uniform -> weight = uniform [-sqrt(6) / sqrt(ni+no) , sqrt(6) / sqrt(ni+ no)]
- He Uniform -> weight = uniform [-sqrt(6 / ni) , sqrt(6 / ni)]
- He Normal -> weight = normal (0 , sqrt(2 / ni))

Where, ni = number of incoming weights at a neuron .

no = number of outgoing weights from a neuron.

Q42. What will happen if we initialize all neurons with same weight?

A42. If we initialize all the neurons with the same weight than we will be sending the same inputs to all the neurons in the hidden layers and the same info will be passed on and the neural network will propagate with the same values again and again. This type of neural network will be useless.

Q43. What are the common data structures used in neural network?

A43. Common data structures used in neural network are:

- Linked List
- Binary Search Tree or Binary Tree
- Heap
- Set
- Graphs
- Hashing

Q44. What are advantages of using neural network?

A44. In neural network the information gets stored in the whole network. It learns from itself and does not require extra input data. They are well known for performing multiple tasks in parallel without affecting the performance of the neural network.

Q45. What are disadvantages of neural network?

A45. When we have a deep neural network i.e. a neural network with many layers, we also have many weights and bias values. So, our ann model in that case tries to over fit.

Q46. What are prerequisite for learning neural network?

A46. To learn and understand how a neural network works you first need to cover the topics listed below:

- Calculus - As there are various formulas and graphs which you need to understand in order to implement neural network.

- Logistic Regression – As most of the neural networks are simultaneously working logistic regressions.
- Coding – Your coding area must be clear in order to implement neural network.

Q47. Most used activation Function?

A47. The most used activation function is the leaky Relu function and softmax function. As they deal with mostly all the limitations or disadvantages in the other activation function.

Q48. Difference between artificial intelligence, machine learning, neural networks ?

A48. Artificial intelligence-It was originated in 1940s. It is basically a process which enables machine to behave like a human being.

Machine learning - It was originated in 1950s. It is basically a study where we use statistical data to make machine work better as it gains experience. It is a part of artificial intelligence.

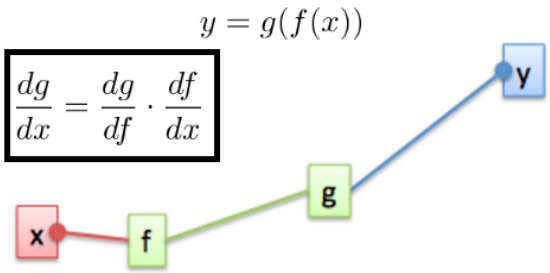
Neural networks -It was originated in 1960s. It is basically imitates the neural network of a human being. It is a part of machine learning.

Q49. What is Vanishing Gradient Problem?

A49. This problem basically arised in the early time of the neural network and is basically due to the sigmoid activation function. As in earlier times, there was only sigmoid activation function discovered. So, what basically happened is sigmoid converts all the values from 0 to 1. But the derivative value of the sigmoid function ranges between “0” to “0.25”. As the layers will be increasing, the value of derivative gets on getting smaller. And a point will come when the new weight will be almost equal to the old weight.

Q50. How chain rule helps in backward propagation?

A50. You can think of a chain rule as a system of traders. From industry to the market.



Q51. Why do we use dropout in neural network?

A51. When we have a deep neural network i.e. neural network with many layers. We will also be having many weights and bias values and due to which our ann model tries to overfit. So, in order to remove/avoid overfitting we use the techniques like dropout.

Q52.What is Dropout in neural network?

A52. So when we have a deep neural network, we assign every layer a dropout value. A dropout value is like a probability of neurons we are choosing. It lies between 0 and 1. So we assign a dropout value for each and every layer in the neural network. When we assign the dropout value than that much neurons will be deactivated and will not be considered in neural network in that very propagation. It chooses random neurons from every layer. It completes one propagation with the activated neurons and the weights are adjusted. Now, in next propagation it will randomly choose neurons.

(Suppose our first layer has 10 neurons and we chose a dropout value (P) as “0.3”. So, 3 neurons randomly will be deactivated.)

Code for initializing neural network:

About the data: We have a file named "Churn_Modelling.csv" . This data set contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer.

Code:

#Importing Libraries

```
import pandas as pd
import keras
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LeakyReLU,PReLU,ELU
from keras.layers import Dropout
```

#Importing Dataset

```
df = pd.read_csvc( "Churn_Modelling.csv" )
df.head()
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10

#Splitting the Dataset into Dependent and independent variable

```
X = df.iloc[:,3:-1]
y = df.iloc[:,-1]
```

#Handling Categorical variables (There are two features namely Gender and Geography which are having categorical values. So in order to remove it, we will apply one hot encoding to them.)

```
Geography = pd.get_dummies(X["Geography"], drop_first=True)

gender = pd.get_dummies ( X["Gender"] , drop_first=True)
```

#Adding the new features to the data frame (Here our data frame is X)

```
X = pd.concat ([ X, geography, gender], axis=1)
```

#Removing the categorical features column from the dataset

```
X = X.drop ([ "Geography" , "Gender"],axis=1)
X.head()
```

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Germany	Spain	Male
0	619	42	2	0.00	1	1	1	101348.88	0	0	0
1	608	41	1	83807.86	1	0	1	112542.58	0	1	0
2	502	42	8	159660.80	3	1	0	113931.57	0	0	0
3	699	39	1	0.00	2	0	0	93826.63	0	0	0
4	850	43	2	125510.82	1	1	1	79084.10	0	1	0

#Splitting the data into training and testing

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X , y , test_size=0.2 ,
random_state=0)
```

#Applying Feature Scaling

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
```

```
X_train = sc.fit_transform ( X_train )
X_test = sc.fit_transform ( X_test )
```

#Initializing the model

```
classifier=Sequential()
```

#Adding input layer and first hidden layer

```
classifier.add(Dense( units = 6 , activation = "relu" , input_dim = 11 , kernel_initializer = "he_uniform"))
```

#Adding second hidden layer

```
classifier.add(Dense(units=6,activation="relu",kernel_initializer="he_uniform"))
```

#Adding output layer

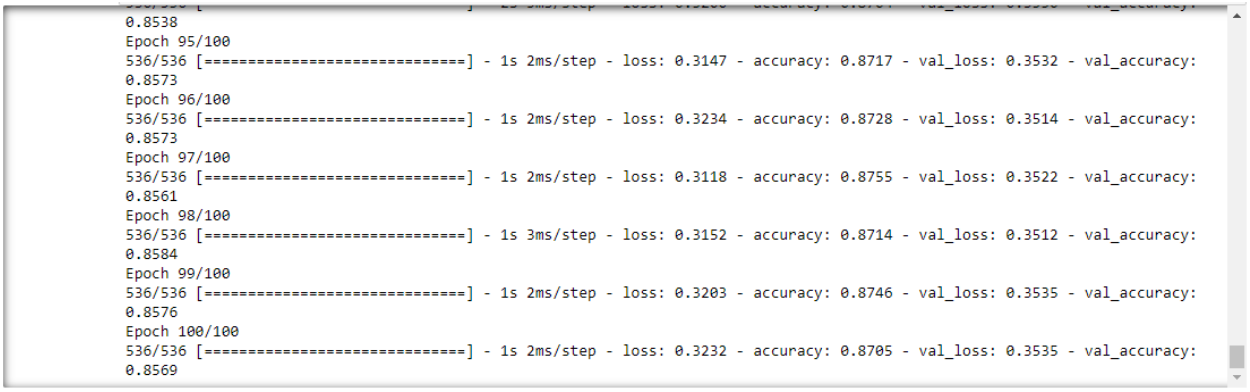
```
classifier.add(Dense(units=1,activation="sigmoid",kernel_initializer="glorot_uniform"))
```

#Compiling the ann model

```
classifier.compile(optimizer="adam",loss="binary_crossentropy",metrics=["accuracy"])
```

#Training the ann model on the training set

model_history = classifier.fit(X_train , y_train , validation_split = 0.33 , batch_size = 10 , epochs=100)



#Predicting the test set results

y_pred = classifier.predict(X_test)
y_pred = (y_pred>0.5)

#Making confusion matrix

from sklearn.metrics import confusion_matrix
cm = confusion_matrix (y_test , y_pred)

#Calculating accuracy score

from sklearn.metrics import accuracy_score
score=accuracy_score(y_pred,y_test)
print(score)

```
0.8625
```

Now, there may rise a question in your mind. Can we apply hyper parameter tuning to select best parameters for our model to perform.

Here is what we will do:

#Importing Libraries

```
Import pandas as pd
from keras.wrappers.scikit_learn import KerasClassifier
from sklearn.model_selection import GridSearchCV
from keras.models import Sequential
from keras.layers import Dense, Activation, Embedding, Flatten, LeakyReLU,
BatchNormalization, Dropout
from keras.activations import relu, sigmoid
```

#Importing Dataset

```
df = pd.read_csv("Churn_Modelling.csv")
df.head()
```

#Splitting the Dataset into independent and dependent dataset

```
X=df.iloc[:,3:-1]
y=df.iloc[:,-1]
```

#Handling Categorical variables (There are two features namely Gender and Geography which are having categorical values. So in order to remove it, we will apply one hot encoding to them.)

```
Geography = pd.get_dummies(X["Geography"], drop_first=True)
gender = pd.get_dummies ( X["Gender"] , drop_first=True)
```

#Adding the new features to the data frame (Here our data frame is X)

```
X = pd.concat ([ X, geography, gender], axis=1)
```

#Removing the categorical features column from the dataset

```
X = X.drop ([ "Geography" , "Gender"],axis=1)
X.head()
```

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Germany	Spain	Male
0	619	42	2	0.00	1	1	1	101348.88	0	0	0
1	608	41	1	83807.86	1	0	1	112542.58	0	1	0
2	502	42	8	159660.80	3	1	0	113931.57	0	0	0
3	699	39	1	0.00	2	0	0	93826.63	0	0	0
4	850	43	2	125510.82	1	1	1	79084.10	0	1	0

#Splitting dataset into train and test set

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X , y , test_size=0.2 ,
random_state=0)
```

#Applying Feature Scaling

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform ( X_train )
X_test = sc.fit_transform ( X_test )
```

#Creating the model with some hyperparameters

```
def create_model(layers, activation):
    model = Sequential()
```

```
    for i, nodes in enumerate(layers):
        if i==0:
            model.add(Dense(nodes,input_dim=X_train.shape[1]))
            model.add(Activation(activation))
            model.add(Dropout(0.3))
        else:
            model.add(Dense(nodes))
            model.add(Activation(activation))
            model.add(Dropout(0.3))

    model.add(Dense(units = 1, kernel_initializer= 'glorot_uniform', activation =
'sigmoid'))
    model.compile(optimizer='adam',
loss='binary_crossentropy',metrics=['accuracy'])
    return model
model = KerasClassifier(build_fn=create_model, verbose=0)
```

#Taking different values and creating a dictionary of hyperparameters

```
layers = [(20), (40, 20), (45, 30, 15)]

activations = ['sigmoid', 'relu']

param_grid = dict(layers=layers, activation=activations, batch_size = [128, 256],
epochs=[30])

grid = GridSearchCV(estimator=model, param_grid=param_grid,cv=5)
```

#Fitting the Dataset into the model

```
grid_result = grid.fit(X_train, y_train)
```

#Best parameters

[grid_result.best_score_,grid_result.best_params_]

```
Out[17]: [0.8567499995231629,
          {'activation': 'relu',
           'batch_size': 128,
           'epochs': 30,
           'layers': (45, 30, 15)}]
```

So, here we got our best parameters to use in order to get the maximum accuracy.