The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Indus Insights

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the difference between Correlation and Regression?

The primary objective of correlation is to measure the degree of linear association between two variables. Here, the dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. Here, the dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be

in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is the difference between R squared and Adjusted R-squared.

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in

the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-

value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

Q: What is the K-Means algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, firstly we select the value of K (the number of clusters) we want. Next, we select the random K points, the centroids (may or may not be from the datasets). Then, assign each data point to their closest centroid and calculate variance and place a new centroid of each cluster. Then reassign each data point to the new closest centroid, if any reassignment took place, we again calculate variance and place the new centroid of each cluster otherwise finish.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better) and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

EXL JPMC Fractal Analytics Mastercard Innovaccer

Tiger Analytics

Q: Explain decision tree algorithm?

Decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: What is the role of C (Regularization) in SVM?

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable. For large values of C, the penalty for misclassifying points is very high, so the decision boundary

will perfectly separate the data if possible. That is why SVM is an example of a large-margin classifier.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load program using smaller initial programs, so in random forest, we divide the dataset between different decision tree and aggregation means to collect, similarly in random forest we are collecting the answers from all the decision trees and provide the majority as a answer.

Q: Explain Type 1 and Type 2 errors.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is regularization and when does it come into play in Machine Learning?

Regularisation is a technique that is used to tackle the problem of the overfitting of the model. When a very complex model is implemented on the training data, it overfits. Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalized and are small in magnitude. L1(lasso regression):

The absolute values of the coefficients are added to the cost function. By reducing the sum of absolute values of the coefficients, what Lasso Regularization (L1 Norm) does is to reduce the number of features in the model altogether to predict the target variable. Basically, it reduces the number of features. L2(Ridge regression): The squares of the coefficients are added to the cost function. It doesn't necessarily reduce the number of features per se, but rather reduces the magnitude/impact that each feature has on the model by reducing the coefficient value. Basically, it reduces the quality of features. It can improve accuracy but is not helpful in selecting variables that are more important as the penalty will shrink all the coefficients. It is also used to tackle multicollinearity.

Q: When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Can the cost function used in linear regression work in logistic regression?

The cost function used in linear regression cannot work with logistic regression. In linear regression, we used the squared error mechanism. Unfortunately for logistic regression, such a cost function produces a nonconvex space that is not ideal for optimization as there will exist many local optima on which our optimization algorithm might prematurely converge before finding the true minimum.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What do you mean by dummy variable and dummy variable trap?

In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity (you will learn more about it later, for now

just remember there is an assumption in the Classical linear regression model that their explanatory variables should be linearly independent of each other, which is violated here). So, to solve this dummy variable trap, the number of dummy variables introduced must be one less than the categories of that variable else you will fall into the dummy variable trap.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - ['Navin', 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - ('Diwakar', 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Difference between sort() and sorted()?

The only difference is sort() function will modify the list it is called on with zero extra space and sorted() function will create a new list containing a sorted version of the list it is given.

Q: What is the difference between R squared and Adjusted R-squared.

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of

how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Q: How do forward propagation and backpropagation work in deep learning?

In short forward propagation is for predicting the output from the model and backpropagation in the minimization of cost function by adjusting the weights in the neural network using Gradient descent. Forward propagation: The inputs are

provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation. Backpropagation is a technique to improve the performance of the network. In neural networks, if the estimated output is far away from the actual output (which means high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. It works by determining the loss (or error) at the output and then propagating it back into the network. The weights are updated to minimize the error resulting from each neuron for this we have to determine the gradient (Derivatives) of each node w.r.t. the final output.

Q: What will happen if the Learning Rate is set too low or too high?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

Q: What is a sigmoid function?

The sigmoid activation function is also called the logistic function. It takes a real-valued number and “squashes” it into a range between 0 and 1, which makes it a very good choice for binary classification. You can classify the output as 0 if it is less than 0.5 and classify it as 1 if the output is more than 0.5. Sigmoid is used in the output layer while making binary predictions. Softmax is used in the output layer while making multi-class predictions. Problems while using the sigmoid function: a) Vanishing gradient problem: A very undesirable property of the sigmoid neuron is that when the neuron’s

activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. We keep on adding more and more Hidden layers in The model, the learning speed of the next hidden layers in the model keeps on getting faster and faster. Now when we do Back-propagation i.e moving backward in the Network and calculating gradients of loss(Error) with respect to the weights, the gradients tend to get smaller and smaller as we keep on moving backward in the Network. This means that the neurons in the Earlier layers learn very slowly as compared to the neurons in the later layers in the Hierarchy. The Earlier layers in the network are slowest to train. The Training process takes too long and the Prediction Accuracy of the Model will decrease. b) Secondly, its output isn't zero-centered. It makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization harder. c) Sigmoids have slow convergence.

Q: Why is random initialization important in NN? Why is it not important in Logistic regression?

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector. If we initialize all the weights with zeros in NN it won't work (initializing bias with zero is ok). If initialized with zeros, then all hidden units would become completely identical (symmetric) and hence compute the exact same function in every iteration. On each gradient descent iteration, all the hidden units will always update in the same way.

Facebook

LinkedIn

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

ShareChat

Q: What is the difference between Correlation and Regression?

In Correlation, the primary objective is to measure the degree of linear association between two variables. The dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. The dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: [True or False] We can just impute the mean for any missing data. It won't affect results and improves power.

False, mean imputation is not a good imputation technique. It does improve power, but your results will be so biased, the improved power won't help much. Sure, your results might be significant, but they're the wrong results!

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a) We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: Explain the decision tree algorithm?

A Decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load the program using smaller initial programs, so in a random forest, we divide the dataset between different decision trees and aggregation means to collect, similarly in the random forest, we are collecting the answers from all the decision trees and provide the majority as an answer.

Q: Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm.

The main reason to use an SVM instead is that the problem might not be linearly separable. In that case, we will have to use an SVM with a non-linear kernel (e.g. RBF). Another related reason to use SVMs is if you are in a higher-

dimensional space. For example, SVMs have been reported to work better for text classification.

Q: What is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Q: What is the difference between Batch Gradient Descent and Stochastic Gradient Descent?

Batch Gradient Descent: The batch gradient computes the gradient using the entire dataset. It takes time to converge because the volume of data is huge, and weights update slowly. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. Stochastic Gradient Descent: The stochastic gradient computes the gradient of a single sample at a time. It converges much faster than the batch gradient because it updates weight more frequently.

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Capgemini

Q: What are the different types of Learning/ Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables: a) Supervised learning: [Target is present]: The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data. Regression Algorithm: The target variable is continuous. Eg: Linear Regression, polynomial Regression, quadratic Regression, decision tree, random forest. Classification Algorithm: The target variable is categorical. Eg: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Bagging and Bagging algorithms, etc. b) Unsupervised learning: It [Target is absent]: The users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data. Clustering Algorithms: K-means clustering and hierarchical clustering. c) Reinforcement Learning: The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

Q: Difference between regression and classification algorithm?

Both Regression and classification are supervised learning algorithms. The difference is: Regression learns from the Labelled Datasets and is then able to predict a continuous-valued (numerical) output for the new data given to the algorithm. In Classification, the algorithm needs to map the new data that is obtained to any one of the 2 classes or it can have more than 2 outputs that we have in our dataset. The classes need to be mapped to either 1 or 0 which in real-life translates to ‘Yes’ or ‘No’, ‘Rains’ or ‘Does Not Rain’, and so forth. The output will be either one of the classes and not a number as it was in Regression.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the difference between python arrays and lists?

The major difference is arrays can hold only a single data type element whereas lists can hold any data type elements. Another difference is lists are generally used for smaller dataset and arrays work more efficiently with the larger dataset.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded

categorical data column and then splits it into multiple columns which are called dummy variables.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null

hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test

dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observation within a cluster and maximize the dissimilarity between clusters.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Quantiphi

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Difference between sort() and sorted()?

The only difference is sort() function will modify the list it is called on with zero extra space and sorted() function will create a new list containing a sorted version of the list it is given.

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job

satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What do you mean by dummy variable and dummy variable trap?

In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity (you will learn more about it later, for now just remember there is an assumption in the Classical linear regression model that their explanatory variables should be linearly independent of each other, which is violated here). So, to solve this dummy variable trap, the number of dummy variables introduced must be one less than the categories of that variable else you will fall into the dummy variable trap.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s)

and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is the difference between R squared and Adjusted R-squared.

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Can the cost function used in linear regression work in logistic regression?

The cost function used in linear regression cannot work with logistic regression. In linear regression, we used the squared error mechanism. Unfortunately for logistic regression, such a cost function produces a nonconvex space that is not ideal for optimization as there will exist many local optima on which our optimization algorithm might prematurely converge before finding the true minimum.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is regularization and when does it come into play in Machine Learning?

Regularisation is a technique that is used to tackle the problem of overfitting of the model. When a very complex model is implemented on the training data, it overfits. Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalized and are small in magnitude. L1(lasso regression): The absolute values of the coefficients are added to the cost function. By reducing the sum of absolute values of the coefficients, what Lasso Regularization (L1 Norm) does is to reduce the number of features in the model altogether to predict the target variable. Basically, it reduces the number of features. L2(Ridge regression): The squares of the coefficients are added to the cost function. It doesn't necessarily reduce the number of features per se, but rather reduces the magnitude/impact that each feature has on the model by reducing the coefficient value. Basically, it reduces the quality of features. It can improve accuracy but is not helpful in selecting variables that are more important as the penalty will shrink all the coefficients. It is also used to tackle multicollinearity.

Q: When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

Q: Explain decision tree algorithm?

Decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is

incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load program using smaller initial programs, so in random forest, we divide the dataset between different decision tree and aggregation means to collect, similarly in random forest we are collecting the answers from all the decision trees and provide the majority as an answer.

Q: What is the role of C (Regularization) in SVM?

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable. For large values of C, the penalty for misclassifying points is very high, so the decision boundary

will perfectly separate the data if possible. That is why SVM is an example of a large-margin classifier.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Q: How do forward propagation and backpropagation work in deep learning?

In short forward propagation is for predicting the output from the model and backpropagation in the minimization of cost function by adjusting the weights in the neural network using Gradient descent. Forward propagation: The inputs are provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation. Backpropagation is a technique to improve the performance of the network. In neural networks, if the estimated output is far away from the actual output (which means high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. It works by determining the loss (or error) at the output and then propagating it back into the network. The weights are updated to minimize the error resulting from each neuron for this we have to determine the gradient (Derivatives) of each node w.r.t. the final output.

Q: What will happen if the Learning Rate is set too low or too high?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

Q: Why is random initialization important in NN? Why is it not important in Logistic regression?

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector. If we initialize all the weights with zeros in NN it won't work (initializing bias with zero is ok). If initialized with zeros, then all hidden units would become completely identical (symmetric) and hence compute the exact same function in every iteration. On each gradient descent iteration, all the hidden units will always update in the same way.

Q: What is a sigmoid function?

The sigmoid activation function is also called the logistic function. It takes a real-valued number and "squashes" it into a range between 0 and 1, which makes it a very good choice for binary classification. You can classify the output as 0 if it is less than 0.5 and classify it as 1 if the output is more than 0.5. Sigmoid is used in the output layer while making binary predictions. Softmax is used in the output layer while making

multi-class predictions. Problems while using the sigmoid function: a) Vanishing gradient problem: A very undesirable property of the sigmoid neuron is that when the neuron's activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. We keep on adding more and more Hidden layers in The model, the learning speed of the next hidden layers in the model keeps on getting faster and faster. Now when we do Back-propagation i.e moving backward in the Network and calculating gradients of loss(Error) with respect to the weights, the gradients tend to get smaller and smaller as we keep on moving backward in the Network. This means that the neurons in the Earlier layers learn very slowly as compared to the neurons in the later layers in the Hierarchy. The Earlier layers in the network are slowest to train. The Training process takes too long and the Prediction Accuracy of the Model will decrease. b) Secondly, its output isn't zero-centered. It makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization harder. c) Sigmoids have slow convergence.

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Oyo Rooms

Q: What is the difference between Correlation and Regression?

In Correlation, the primary objective is to measure the degree of linear association between two variables. The dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. The dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying

logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a) We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: Explain the decision tree algorithm?

A Decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class

with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: [True or False] We can just impute the mean for any missing data. It won't affect results and improves power.

False, mean imputation is not a good imputation technique. It does improve power, but your results will be so biased, the improved power won't help much. Sure, your results might be significant, but they're the wrong results!

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational

variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm.

The main reason to use an SVM instead is that the problem might not be linearly separable. In that case, we will have to use an SVM with a non-linear kernel (e.g. RBF). Another related reason to use SVMs is if you are in a higher-dimensional space. For example, SVMs have been reported to work better for text classification.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load the program using smaller initial programs, so in a random forest, we divide the dataset between different decision trees and aggregation means to collect, similarly in the random forest, we are collecting the

answers from all the decision trees and provide the majority as an answer.

Q: What is the difference between Batch Gradient Descent and Stochastic Gradient Descent?

Batch Gradient Descent: The batch gradient computes the gradient using the entire dataset. It takes time to converge because the volume of data is huge, and weights update slowly. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. Stochastic Gradient Descent: The stochastic gradient computes the gradient of a single sample at a time. It converges much faster than the batch gradient because it updates weight more frequently.

Q: What is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

EXL JPMC Fractal Analytics Mastercard Innovaccer

Axis Bank

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

Q: What is the K-Means algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, firstly we select the value of K (the number of clusters) we want. Next, we select the random K points, the centroids (may or may not be from the datasets). Then, assign each data point to their closest centroid and calculate variance and place a new centroid of each cluster. Then reassign each data point to the new closest centroid, if any reassignment took place, we again calculate variance and place the new centroid of each cluster otherwise finish.

Q: What is the difference between Correlation and Regression?

The primary objective of correlation is to measure the degree of linear association between two variables. Here, the dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. Here, the dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies

that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be

in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null

hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying

logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better) and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

FOLLOW US

Facebook

LinkedIn

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Ernst & young GDS

Q: What are the different types of Learning/ Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables: a) Supervised learning: [Target is present]: The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data. Regression Algorithm: The target variable is continuous. Eg: Linear Regression, polynomial Regression, quadratic Regression, decision tree, random forest. Classification Algorithm: The target variable is categorical. Eg: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Bagging and Bagging algorithms, etc. b) Unsupervised learning: It [Target is absent]: The users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data. Clustering Algorithms: K-means clustering and hierarchical clustering. c) Reinforcement Learning: The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

Q: What is the difference between python arrays and lists?

The major difference is arrays can hold only a single data type element whereas lists can hold any data type elements. Another difference is lists are generally used for smaller dataset and arrays work more efficiently with the larger dataset.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s)

and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Difference between regression and classification algorithm?

Both Regression and classification are supervised learning algorithms. The difference is: Regression learns from the Labelled Datasets and is then able to predict a continuous-

valued (numerical) output for the new data given to the algorithm. In Classification, the algorithm needs to map the new data that is obtained to any one of the 2 classes or it can have more than 2 outputs that we have in our dataset. The classes need to be mapped to either 1 or 0 which in real-life translates to 'Yes' or 'No', 'Rains' or 'Does Not Rain', and so forth. The output will be either one of the classes and not a number as it was in Regression.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observation within a cluster and maximize the dissimilarity between clusters.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null

hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

TCS Research and Innovation Lab

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p-value of the added variable, for the variable to be significant, the p-value should be less than the level of significance.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with

changes in the response. A high P-value is also called an insignificant P-value.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job

satisfaction, and memory, etc are distributed approximately normally.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Difference between sort() and sorted()?

The only difference is sort() function will modify the list it is called on with zero extra space and sorted() function will create a new list containing a sorted version of the list it is given.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded

categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What do you mean by dummy variable and dummy variable trap?

In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity (you will learn more about it later, for now just remember there is an assumption in the Classical linear regression model that their explanatory variables should be linearly independent of each other, which is violated here). So, to solve this dummy variable trap, the number of dummy variables introduced must be one less than the categories of that variable else you will fall into the dummy variable trap.

Q: Can the cost function used in linear regression work in logistic regression?

The cost function used in linear regression cannot work with logistic regression. In linear regression, we used the squared error mechanism. Unfortunately for logistic regression, such a cost function produces a nonconvex space that is not ideal for optimization as there will exist many local optima on which our optimization algorithm might prematurely converge before finding the true minimum.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In the case of a logistic regression model, the decision boundary is a straight line.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: What is regularization and when does it come into play in Machine Learning?

Regularisation is a technique that is used to tackle the problem of overfitting of the model. When a very complex model is implemented on the training data, it overfits. Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalized and are small in magnitude. L1(lasso regression): The absolute values of the coefficients are added to the cost function. By reducing the sum of absolute values of the coefficients, what Lasso Regularization (L1 Norm) does is to reduce the number of features in the model altogether to predict the target variable. Basically, it reduces the number of features. L2(Ridge regression): The squares of the coefficients are added to the cost function. It doesn't necessarily reduce the number of features per se, but rather reduces the magnitude/impact that each feature has on the model by reducing the coefficient value. Basically, it reduces the quality

of features. It can improve accuracy but is not helpful in selecting variables that are more important as the penalty will shrink all the coefficients. It is also used to tackle multicollinearity.

Q: When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: Explain the decision tree algorithm?

A decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with

decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load program using smaller initial programs, so in random forest, we divide the dataset between different decision tree and aggregation means to collect, similarly in random forest we are collecting the answers from all the decision trees and provide the majority as a answer.

Q: What is the role of C (Regularization) in SVM?

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable. For large values of C, the penalty for misclassifying points is very high, so the decision boundary

will perfectly separate the data if possible. That is why SVM is an example of a large-margin classifier.

Q: How does forward propagation and backpropagation work in deep learning?

In short forward propagation is for predicting the output from the model and backpropagation in the minimization of cost function by adjusting the weights in the neural network using Gradient descent. Forward propagation: The inputs are provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation. Backpropagation is a technique to improve the performance of the network. In neural networks, if the estimated output is far away from the actual output (which means high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. It works by determining the loss (or error) at the output and then propagating it back into the network. The weights are updated to minimize the error resulting from each neuron for this we have to determine the gradient (Derivatives) of each node w.r.t. the final output.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large-scale neural network.

Q: Why is random initialization important in NN? Why is it not important in Logistic regression?

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector. If we initialize all the weights with zeros in NN it won't work (initializing bias with zero is ok). If initialized with zeros, then all hidden units would become completely identical (symmetric) and hence compute the exact same function in every iteration. On each gradient descent iteration, all the hidden units will always update in the same way.

Q: What is a sigmoid function?

The sigmoid activation function is also called the logistic function. It takes a real-valued number and "squashes" it into a range between 0 and 1, which makes it a very good choice for binary classification. You can classify the output as 0 if it is less than 0.5 and classify it as 1 if the output is more than 0.5. Sigmoid is used in the output layer while making binary predictions. Softmax is used in the output layer while making multi-class predictions. Problems while using the sigmoid function: a) Vanishing gradient problem: A very undesirable property of the sigmoid neuron is that when the neuron's activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. We keep on adding more and more Hidden layers in The model, the learning speed of the next hidden layers in the model keeps on getting faster and faster. Now when we do Back-propagation i.e moving backward in the Network and calculating gradients of loss(Error) with respect to the weights, the gradients tend to get smaller and smaller as we keep on moving backward in the Network. This means that the neurons in the Earlier

layers learn very slowly as compared to the neurons in the later layers in the Hierarchy. The Earlier layers in the network are slowest to train. The Training process takes too long and the Prediction Accuracy of the Model will decrease. b) Secondly, its output isn't zero-centered. It makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization harder. c) Sigmoids have slow convergence.

Q: What will happen if the Learning Rate is set too low or too high?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

MX Player

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is the difference between Correlation and Regression?

In Correlation, the primary objective is to measure the degree of linear association between two variables. The dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. The dependent variable is assumed to be random while

independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: [True or False] We can just impute the mean for any missing data. It won't affect results and improves power.

False, mean imputation is not a good imputation technique. It does improve power, but your results will be so biased, the improved power won't help much. Sure, your results might be significant, but they're the wrong results!

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the

response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a) We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: Explain the decision tree algorithm?

A Decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm.

The main reason to use an SVM instead is that the problem might not be linearly separable. In that case, we will have to use an SVM with a non-linear kernel (e.g. RBF). Another related reason to use SVMs is if you are in a higher-dimensional space. For example, SVMs have been reported to work better for text classification.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load the program using smaller initial programs, so in a random forest, we divide the dataset between different decision trees and aggregation means to collect, similarly in the random forest, we are collecting the

answers from all the decision trees and provide the majority as an answer.

Q: What is the difference between Batch Gradient Descent and Stochastic Gradient Descent?

Batch Gradient Descent: The batch gradient computes the gradient using the entire dataset. It takes time to converge because the volume of data is huge, and weights update slowly. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. Stochastic Gradient Descent: The stochastic gradient computes the gradient of a single sample at a time. It converges much faster than the batch gradient because it updates weight more frequently.

Q: What is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large-scale neural network.

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

Data Science Interview Question

Recommended companies

EXL › JPMC › Fractal Analytics › Mastercard › Innovaccer ›

Citi Bank

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is the difference between Correlation and Regression?

The primary objective of correlation is to measure the degree of linear association between two variables. Here, the dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. Here, the dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: What is the difference between R squared and Adjusted R-squared.

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - ['Navin', 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - ('Diwakar', 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text

data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of

making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: What is the K-Means algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, firstly we select the value of K (the number of clusters) we want. Next, we select the random K points, the centroids (may or may not be from the datasets). Then, assign each data point to their closest centroid and calculate variance and place a new centroid of each cluster. Then reassign each data point to the new closest centroid, if any reassignment took place, we again calculate variance and place the new centroid of each cluster otherwise finish.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better) and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Cred

Q: Difference between regression and classification algorithm?

Both Regression and classification are supervised learning algorithms. The difference is: Regression learns from the Labelled Datasets and is then able to predict a continuous-valued (numerical) output for the new data given to the algorithm. In Classification, the algorithm needs to map the new data that is obtained to any one of the 2 classes or it can have more than 2 outputs that we have in our dataset. The classes need to be mapped to either 1 or 0 which in real-life translates to 'Yes' or 'No', 'Rains' or 'Does Not Rain', and so forth. The output will be either one of the classes and not a number as it was in Regression.

Q: What is the difference between python arrays and lists?

The major difference is arrays can hold only a single data type element whereas lists can hold any data type elements. Another difference is lists are generally used for smaller dataset and arrays work more efficiently with the larger dataset.

Q: What are the different types of Learning/ Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables: a) Supervised learning: [Target is present]: The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data. Regression Algorithm: The target variable is continuous. Eg: Linear Regression, polynomial Regression, quadratic Regression, decision tree, random forest. Classification Algorithm: The target variable is categorical. Eg: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Bagging and Bagging algorithms, etc. b) Unsupervised learning: It [Target is absent]: The users do not need to supervise the model. Instead, it allows the model to

work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data. Clustering Algorithms: K-means clustering and hierarchical clustering. c) Reinforcement Learning: The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: tate the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity:

it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p-value of the added variable, for the variable to be significant, the p-value should be less than the level of significance.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observation within a cluster and maximize the dissimilarity between clusters.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

EXL JPMC Fractal Analytics Mastercard Innovaccer

Accenture India

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Difference between sort() and sorted()?

The only difference is sort() function will modify the list it is called on with zero extra space and sorted() function will

create a new list containing a sorted version of the list it is given.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What do you mean by dummy variable and dummy variable trap?

In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity (you will learn more about it later, for now just remember there is an assumption in the Classical linear regression model that their explanatory variables should be linearly independent of each other, which is violated here). So, to solve this dummy variable trap, the number of dummy variables introduced must be one less than the categories of that variable else you will fall into the dummy variable trap.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R

square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: Can the cost function used in linear regression work in logistic regression?

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data

should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: Explain the decision tree algorithm?

A decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with

decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: What is regularization and when does it come into play in Machine Learning?

Regularisation is a technique that is used to tackle the problem of the overfitting of the model. When a very complex model is implemented on the training data, it overfits. Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalized and are small in magnitude. L1(lasso regression): The absolute values of the coefficients are added to the cost function. By reducing the sum of absolute values of the coefficients, what Lasso Regularization (L1 Norm) does is to reduce the number of features in the model altogether to predict the target variable. Basically, it reduces the number of features. L2(Ridge regression): The squares of the coefficients are added to the cost function. It doesn't necessarily reduce the number of features per se, but rather reduces the magnitude/impact that each feature has on the model by reducing the coefficient value. Basically, it reduces the quality of features. It can improve accuracy but is not helpful in selecting variables that are more important as the penalty will shrink all the coefficients. It is also used to tackle multicollinearity.

Q: When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load program using smaller initial programs, so in random forest, we divide the dataset between different decision tree and aggregation means to collect, similarly in random forest we are collecting the answers from all the decision trees and provide the majority as a answer.

Q: What is the role of C (Regularization) in SVM?

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable. For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible. That is why SVM is an example of a large-margin classifier.

Q: How do forward propagation and backpropagation work in deep learning?

In short forward propagation is for predicting the output from the model and backpropagation in the minimization of

cost function by adjusting the weights in the neural network using Gradient descent. Forward propagation: The inputs are provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation. Backpropagation is a technique to improve the performance of the network. In neural networks, if the estimated output is far away from the actual output (which means high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. It works by determining the loss (or error) at the output and then propagating it back into the network. The weights are updated to minimize the error resulting from each neuron for this we have to determine the gradient (Derivatives) of each node w.r.t. the final output.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large-scale neural network.

Q: What will happen if the Learning Rate is set too low or too high?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge

(model can give a good output) or even diverge (data is too chaotic for the network to train).

Q: Why is random initialization important in NN? Why is it not important in Logistic regression?

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector. If we initialize all the weights with zeros in NN it won't work (initializing bias with zero is ok). If initialized with zeros, then all hidden units would become completely identical (symmetric) and hence compute the exact same function in every iteration. On each gradient descent iteration, all the hidden units will always update in the same way.

Q: What is a sigmoid function?

The sigmoid activation function is also called the logistic function. It takes a real-valued number and "squashes" it into a range between 0 and 1, which makes it a very good choice for binary classification. You can classify the output as 0 if it is less than 0.5 and classify it as 1 if the output is more than 0.5. Sigmoid is used in the output layer while making binary predictions. Softmax is used in the output layer while making multi-class predictions. Problems while using the sigmoid function: a) Vanishing gradient problem: A very undesirable property of the sigmoid neuron is that when the neuron's activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. We keep on adding more and more Hidden layers in The model, the learning speed of the next hidden layers in the model keeps on getting faster and faster. Now when we do Back-propagation i.e moving backward in the Network and calculating gradients of loss(Error) with respect to the weights, the gradients tend to

get smaller and smaller as we keep on moving backward in the Network. This means that the neurons in the Earlier layers learn very slowly as compared to the neurons in the later layers in the Hierarchy. The Earlier layers in the network are slowest to train. The Training process takes too long and the Prediction Accuracy of the Model will decrease. b) Secondly, its output isn't zero-centered. It makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization harder. c) Sigmoids have slow convergence.

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

American Express

Q: What is the difference between Correlation and Regression?

The primary objective of correlation is to measure the degree of linear association between two variables. Here, the dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. Here, the dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test

dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the K-Means algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, firstly we select the value of K (the number of clusters) we want. Next, we select the random K points, the centroids (may or may not be from the datasets). Then, assign each data point to their closest centroid and calculate variance and place a new centroid of each cluster. Then reassign each data point to the new closest centroid, if any reassignment took place, we again calculate variance and place the new centroid of each cluster otherwise finish.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a

convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Formula: R^2 = ESS/TSS = 1-RSS/TSS. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Formula: Adjusted R^2 = 1-((1-R2)(n-1)/n-k-1). But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded

categorical data column and then splits it into multiple columns which are called dummy variables.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - ['Navin', 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - ('Diwakar', 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better) and an output layer. Each sheet contains neurons called "nodes," performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Facebook

LinkedIn

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Goldman Sachs

Q: What is the difference between correlation and causation?

Causation explicitly applies to cases where action A causes outcome B. On the other hand, correlation is simply a relationship, here action A relates to action B—but one event doesn't necessarily cause the other event to happen. While causation and correlation can exist at the same time, correlation does not imply causation. Eg: more sleep will cause you to perform better at work. Or, more cardio will cause you to lose your belly fat.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. R^2=ESS/TSS=1-RSS/TSS. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. But an independent variable that has a

correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic. Adjusted R^2=1-((1-R^2)(n-1)/n-k-1)

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: Explain Ordinary Least Squares Regression in brief.

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression is used in a bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression can also be used in multivariate models in which there are two or more independent variables.

Q: Why do we square the error instead of using modulus?

It's true that one could choose to use the absolute error instead of the squared error. In fact, the absolute error is often closer to what we want when making predictions from our model. But, we want to penalize those predicted values which are contributing the maximum error. Moreover,

looking a little deeper, the squared error is everywhere differentiable, while the absolute error is not (its derivative is undefined at 0). This makes the squared error more amenable to the techniques of mathematical optimization. To optimize the squared error, we can just set its derivative equal to 0 and solve. To optimize the absolute error often requires more complex techniques. Actually, we find the Root Mean Squared Error so that the unit of RMSE and the dependent variable are equal.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3:

The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: Why is it assumed that the error term is normally distributed?

This can be justified by Central Limit Theorem. It can be shown that if there are a large number of independent and identically distributed random variables, then, with a few exceptions, the distribution of their sum tends to a normal distribution as the number of such variables increases indefinitely.

Q: Define Multicollinearity and why does the linear regression model assume that there should be no multicollinearity among the explanatory variables?

Multicollinearity means the existence of a “perfect,” or exact, linear relationship among some or all explanatory variables of a regression model. It does not rule out non linear relationships among them. In the case of perfect multicollinearity one cannot get a unique solution for the individual regression coefficients. But one can get a unique solution for linear combinations of these coefficients. And it can increase the variance of the regression coefficients,

making them unstable and difficult to interpret; that is why, linear regression assumes no multicollinearity.

Q: How to measure Multicollinearity?

We can check for the presence of Multicollinearity using Variance Inflation Factor(VIF). VIF shows how the variance is inflated by the presence of multicollinearity, which means how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity. Some other methods are also there, but VIF is the most important of them all.

Q: How to tackle Multicollinearity?

To deal with Multicollinearity Try any one of the following methods:- a) Remove highly correlated predictors from the model. If there exist two or more factors with high VIF, remove one of them since they supply redundant information. b) Linearly combine the independent variables, such as adding them together. c) Perform an analysis designed for highly correlated variables, such as principal components analysis. PCA reduces the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example -

('Diwakar', 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Differentiate between Mean Square Error and Mean Absolute Error.

Mean Square Error(MSE) is the sum of the squared distances between our target variable and predicted variables. Advantage: The MSE is great for ensuring that our trained model has no outlier predictions with huge errors since the MSE puts larger weight on these errors due to the squaring part of the function Disadvantage: If our model makes a single very bad prediction, the squaring part of the function magnifies the error. Mean Absolute Error(MAE) is the sum of absolute differences between our target and predicted variables. So it measures the average magnitude of errors in a set of predictions, without considering their directions. Advantage: The beauty of the MAE is that its advantage directly covers the MSE disadvantage. Since we are taking

the absolute value, all of the errors will be weighted on the same linear scale. Thus, unlike the MSE, we won't be putting too much weight on our outliers and our loss function provides a generic and even measure of how well our model is performing. Disadvantage: If we do in fact care about the outlier predictions of our model, then the MAE won't be as effective. The large errors coming from the outliers end up being weighted the exact same as lower errors. This might result in our model being great most of the time, but making a few very poor predictions every-so-often

Q: Which evaluation technique should you prefer to use for data having a lot of outliers in it?

Mean Absolute Error(MAE) is preferable to use for data having too many outliers in it because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and starts penalizing the outliers by squaring the residuals.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This

means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect, when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large-scale neural network.

Q: What is Gradient descent?

Gradient Descent is an optimal algorithm to minimize the cost function or to minimize an error. The aim is to find the local-global minima of a function. This determines the direction the model should take to reduce the error.

Facebook

LinkedIn

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

Innovaccer

Q: What are the different types of Learning/ Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables: a) Supervised learning: [Target is present]: The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data. Regression Algorithm: The target variable is continuous. Eg: Linear Regression, polynomial Regression, quadratic Regression, decision tree, random forest. Classification Algorithm: The target variable is categorical. Eg: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Bagging and Bagging algorithms, etc. b) Unsupervised learning: It [Target is absent]: The users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data. Clustering Algorithms: K-means clustering and hierarchical clustering. c) Reinforcement Learning: The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

Q: Difference between regression and classification algorithm?

Both Regression and classification are supervised learning algorithms. The difference is: Regression learns from the Labelled Datasets and is then able to predict a continuous-valued (numerical) output for the new data given to the algorithm. In Classification, the algorithm needs to map the new data that is obtained to any one of the 2 classes or it can have more than 2 outputs that we have in our dataset. The classes need to be mapped to either 1 or 0 which in real-life translates to ‘Yes’ or ‘No’, ‘Rains’ or ‘Does Not Rain’, and so forth. The output will be either one of the classes and not a number as it was in Regression.

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set, and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the difference between python arrays and lists?

The major difference is arrays can hold only a single data type element whereas lists can hold any data type elements. Another difference is lists are generally used for smaller dataset and arrays work more efficiently with the larger dataset.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded

categorical data column and then splits it into multiple columns which are called dummy variables.

Q: State the assumption of a Classical Linear Regression model.

Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variables should be linearly independent of each other.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p value of the added variable, for the variable to be significant, p value should be less than the level of significance.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null

hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a)We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test

dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the goal of clustering analysis?

The goal of clustering analysis is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

EXL | JPMC | Fractal Analytics | Mastercard | Innovaccer

Mastercard

Q: What is the difference between Correlation and Regression?

In Correlation, the primary objective is to measure the degree of linear association between two variables. The dependent and independent variables are assumed to be random. In regression, we try to estimate or predict the average value of one var on the basis of fixed values of other variables. The dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random).

Q: What is the normal distribution and why is it so important?

The distribution having the mean, mode, and median all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right, and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: [True or False] We can just impute the mean for any missing data. It won't affect results and improves power.

False, mean imputation is not a good imputation technique. It does improve power, but your results will be so biased, the improved power won't help much. Sure, your results might be significant, but they're the wrong results!

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: What is logistic regression?

Logistic Regression is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and the probability of a particular outcome. Example: When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a) We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It's helpful in reducing the error.

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests. Underfitting happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: Explain the decision tree algorithm?

A decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load the program using smaller initial programs, so in a random forest, we divide the dataset between different decision trees and aggregation means to collect, similarly in the random forest, we are collecting the answers from all the decision trees and provide the majority as an answer.

Q: Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm.

The main reason to use an SVM instead is that the problem might not be linearly separable. In that case, we will have to use an SVM with a non-linear kernel (e.g. RBF). Another related reason to use SVMs is if you are in a higher-

dimensional space. For example, SVMs have been reported to work better for text classification.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Q: What is the difference between Batch Gradient Descent and Stochastic Gradient Descent?

Batch Gradient Descent: The batch gradient computes the gradient using the entire dataset. It takes time to converge because the volume of data is huge, and weights update slowly. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. Stochastic Gradient Descent: The stochastic gradient computes the gradient of a single sample at a time. It converges much faster than the batch gradient because it updates weight more frequently.

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

Data Science Interview Question

Recommended companies

EXL › JPMC › Fractal Analytics › Mastercard › Innovaccer ›

Fractal Analytics

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - List, tuple, set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: Difference between sort() and sorted()?

The only difference is sort() function will modify the list it is called on with zero extra space and sorted() function will create a new list containing a sorted version of the list it is given.

Q: What is the normal distribution and why is it so important?

The distribution in which mean, mode, and median are all equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job

satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: What do you mean by dummy variable and dummy variable trap?

In regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity (you will learn more about it later, for now just remember there is an assumption in the Classical linear regression model that their explanatory variables should be linearly independent of each other, which is violated here). So, to solve this dummy variable trap, the number of dummy variables introduced must be one less than the categories of that variable else you will fall into the dummy variable trap.

Q: State the assumption of a Classical Linear Regression model.

The Classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s)

and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow normal distribution; it means zero mean value and constant standard deviation.[It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variable should be linearly independent of each other.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. R^2=ESS/TSS=1-RSS/TSS. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic. Adjusted R2=1-((1-R2)(n-1)/n-k-1)

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p-value of the added variable, for the variable to be significant, the p-value should be less than the level of significance.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Can the cost function used in linear regression work in logistic regression?

The cost function used in linear regression cannot work with logistic regression. In linear regression, we used the squared error mechanism. Unfortunately for logistic regression, such a cost function produces a nonconvex space that is not ideal for optimization as there will exist many local optima on which our optimization algorithm might prematurely converge before finding the true minimum.

Q: Is the decision boundary linear or nonlinear in the case of a logistic regression model?

The decision boundary is a line that separates the target variables into different classes. The decision boundary can either be linear or nonlinear. In case of a logistic regression model, the decision boundary is a straight line.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If we have a high bias error what does it mean? How to treat it?

High bias error means that the model we are using is ignoring all the important trends in the model and the model is underfitting. To reduce underfitting: a) We need to increase the complexity of the model. b) The number of features needs to be increased. Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that the most important signals are found by the model to make effective predictions. Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

Q: What is regularization and when does it come into play in Machine Learning?

Regularisation is a technique that is used to tackle the problem of the overfitting of the model. When a very complex model is implemented on the training data, it overfits. Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalized and are small in magnitude. L1(lasso regression): The absolute values of the coefficients are added to the cost function. By reducing the sum of absolute values of the coefficients, what Lasso Regularization (L1 Norm) does is to reduce the number of features in the model altogether to predict the target variable. Basically, it reduces the number of features. L2(Ridge regression): The squares of the coefficients are added to the cost function. It doesn't necessarily reduce the number of features per se, but rather reduces the magnitude/impact that each feature has on the model by reducing the coefficient value. Basically, it reduces the quality of features. It can improve accuracy but is not helpful in selecting variables that are more important as the penalty will shrink all the coefficients. It is also used to tackle multicollinearity.

Q: When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

Q: Explain the decision tree algorithm?

A decision tree is a supervised learning algorithm, which can be used for both regression and classification, but mostly is used for classification. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Basically, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is

incrementally developed. The final result is a tree with decision nodes and leaf nodes. In this, two measures are important - entropy and information gain.

Q: What is entropy?

Entropy is defined as a measure of impurity of the data. Data with the high disorder can be said to be data with high entropy, and, homogenous (or pure) data can be termed as data with very low entropy. The value of entropy close to zero represents the fact that data is pure. This essentially means that data belongs to one or mostly one class level (the class with a label). A value closer to 1 represents maximum disorder or maximum split. This implies 50-50 or equal splits in the data segment. So, less entropy is better.

Q: Why is random forest also called bootstrap aggregation?

Random forest is also called bootstrap aggregation because bootstrap means to load the program using smaller initial programs, so in a random forest, we divide the dataset between different decision trees and aggregation means to collect, similarly in the random forest, we are collecting the answers from all the decision trees and provide the majority as an answer.

Q: What is the role of C (Regularization) in SVM?

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable. For large values of C, the penalty for misclassifying points is very high, so the decision boundary

will perfectly separate the data if possible. That is why SVM is an example of a large-margin classifier.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large-scale neural network.

Q: How do forward propagation and backpropagation work in deep learning?

In short forward propagation is for predicting the output from the model and backpropagation in the minimization of cost function by adjusting the weights in the neural network using Gradient descent. Forward propagation: The inputs are provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation. Backpropagation is a technique to improve the performance of the network. In neural networks, if the estimated output is far away from the actual output (which means high error), we update the biases and weights based on the error. This weight and bias updating process is known as Back Propagation. It works by determining the loss (or error) at the output and then propagating it back into the network. The weights are updated to minimize the error resulting from each neuron for this we have to determine the gradient (Derivatives) of each node w.r.t. the final output.

Q: What will happen if the Learning Rate is set too low or too high?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

Q: Why is random initialization important in NN? Why is it not important in Logistic regression?

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the values of the weight follow x's distribution and are different from each other if x is not a constant vector. If we initialize all the weights with zeros in NN it won't work (initializing bias with zero is ok). If initialized with zeros, then all hidden units would become completely identical (symmetric) and hence compute the exact same function in every iteration. On each gradient descent iteration, all the hidden units will always update in the same way.

Q: What is a sigmoid function?

The sigmoid activation function is also called the logistic function. It takes a real-valued number and "squashes" it into a range between 0 and 1, which makes it a very good choice for binary classification. You can classify the output as 0 if it is less than 0.5 and classify it as 1 if the output is more than 0.5. Sigmoid is used in the output layer while making binary

predictions. Softmax is used in the output layer while making multi-class predictions.

Q: Explain the problems while using the sigmoid function.

a) Vanishing gradient problem: A very undesirable property of the sigmoid neuron is that when the neuron's activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. We keep on adding more and more Hidden layers in The model, the learning speed of the next hidden layers in the model keeps on getting faster and faster. Now when we do Back-propagation i.e moving backward in the Network and calculating gradients of loss(Error) with respect to the weights, the gradients tend to get smaller and smaller as we keep on moving backward in the Network. This means that the neurons in the Earlier layers learn very slowly as compared to the neurons in the later layers in the Hierarchy. The Earlier layers in the network are slowest to train. The Training process takes too long and the Prediction Accuracy of the Model will decrease. b) Secondly, its output isn't zero-centered. It makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization harder. c) Sigmoids have slow convergence.

Contact Us

contact@thenoncore.com

The Non Core.com

Data Science Interview Question

Recommended companies

- EXL
- JPMC
- Fractal Analytics
- Mastercard
- Innovaccer

JPMC

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple, set, and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the difference between correlation and causation?

Causation explicitly applies to cases where action A causes outcome B. On the other hand, correlation is simply a relationship, here action A relates to action B—but one event doesn’t necessarily cause the other event to happen. While causation and correlation can exist at the same time, correlation does not imply causation. Eg: more sleep will cause you to perform better at work. Or, more cardio will cause you to lose your belly fat.

Q: What is the normal distribution and why is it so important?

The distribution in which mean, mode, and median all are equal, the curve is symmetric at the center (i.e. around the mean, μ), exactly half of the values are to the left of the center and exactly half the values are to the right and the total area under the curve is 1 is called the normal distribution. Normality assumptions in linear regression play a very important role and it is easy for mathematical statisticians to

work with as many kinds of statistical tests can be derived for normal distributions. Many psychological and educational variables like measures of reading ability, introversion, job satisfaction, and memory, etc are distributed approximately normally.

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: State the assumption of a Classical Linear Regression model.

The classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow a normal distribution; it means zero mean value and constant standard deviation. [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error.]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory

variables). A7: No Multicollinearity: The independent variable should be linearly independent of each other.

Q: Why is it assumed that the error term is normally distributed?

This can be justified by the Central Limit Theorem. It can be shown that if there are a large number of independent and identically distributed random variables, then, with a few exceptions, the distribution of their sum tends to a normal distribution as the number of such variables increases indefinitely.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic. R^2=ESS/TSS=1-RSS/TSS. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. But an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic. Adjusted R^2=1-((1-R2)(n-1)/n-k-1)

Q: Define Multicollinearity and why does the linear regression model assume that there should be no multicollinearity among the explanatory variables?

Multicollinearity means the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of

a regression model. It does not rule out nonlinear relationships among them. In the case of perfect multicollinearity, one cannot get a unique solution for the individual regression coefficients. But one can get a unique solution for linear combinations of these coefficients. And it can increase the variance of the regression coefficients, making them unstable and difficult to interpret; that is why linear regression assumes no multicollinearity.

Q: How to measure Multicollinearity?

We can check for the presence of Multicollinearity using Variance Inflation Factor(VIF). VIF shows how the variance is inflated by the presence of multicollinearity, which means how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1. A VIF between 5 and 10 indicates a high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity. Some other methods are also there, but VIF is the most important of them all.

Q: How to tackle Multicollinearity?

To deal with Multicollinearity, try any one of the following methods:- a) Remove highly correlated predictors from the model. If there exist two or more factors with high VIF, remove one of them since they supply redundant information. b) Linearly combine the independent variables, such as adding them together. c) Perform an analysis designed for highly correlated variables, such as principal components analysis. PCA reduces the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

Q: What happens when we add a variable and it increases the R square but decreases the Adjusted R Square?

The variable can be omitted since it holds no predictive power and we should also look at the p-value of the added variable, for the variable to be significant, p-value should be less than the level of significance.

Q: Differentiate between Mean Square Error and Mean Absolute Error.

Mean Square Error(MSE) is the sum of the squared distances between our target variable and predicted variables. Advantage: The MSE is great for ensuring that our trained model has no outlier predictions with huge errors, since the MSE puts larger weight on these errors due to the squaring part of the function. Disadvantage: If our model makes a single very bad prediction, the squaring part of the function magnifies the error. Mean Absolute Error(MAE) is the sum of absolute differences between our target and predicted variables. So it measures the average magnitude of errors in a set of predictions, without considering their directions. Advantage: The beauty of the MAE is that its advantage directly covers the MSE disadvantage. Since we are taking the absolute value, all of the errors will be weighted on the same linear scale. Thus, unlike the MSE, we won't be putting too much weight on our outliers and our loss function provides a generic and even measure of how well our model is performing. Disadvantage: If we do in fact care about the outlier predictions of our model, then the MAE won't be as effective. The large errors coming from the outliers end up being weighted the exact same as lower errors. This might result in our model being great most of the time, but making a few very poor predictions every-so-often

Q: Explain Ordinary Least Squares Regression in brief.

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression is used in a bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression can also be used in multivariate models in which there are two or more independent variables.

Q: Which evaluation technique should you prefer to use for data having a lot of outliers in it?

Mean Absolute Error(MAE) is preferable to use for data having too many outliers in it because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and starts penalizing the outliers by squaring the residuals.

Q: Why do we square the error instead of using modulus?

It’s true that one could choose to use the absolute error instead of the squared error. In fact, the absolute error is often closer to what we want when making predictions from our model. But, we want to penalize those predicted values which are contributing the maximum error. Moreover, looking a little deeper, the squared error is everywhere differentiable, while the absolute error is not (its derivative is undefined at 0). This makes the squared error more amenable to the techniques of mathematical optimization. To optimize the squared error, we can just set its derivative equal to 0 and solve. To optimize the absolute error often requires more complex techniques. Actually, we find the Root

Mean Squared Error so that the unit of RMSE and the dependent variable are equal.

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less

likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model.

Q: What Is the Role of Activation Functions in a Neural Network?

An activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Sigmoid, ReLU, Leaky ReLU, Tanh, and Softmax are examples of activation functions. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model, which can make or break a large scale neural network.

Q: What is Gradient descent?

Gradient Descent is an optimal algorithm to minimize the cost function or to minimize an error. The aim is to find the local-global minima of a function. This determines the direction the model should take to reduce the error.

FOLLOW US

Facebook

LinkedIn

Telegram

PATHS

Data Science

Practice Test

Interview Questions

Contact Us

contact@thenoncore.com

Data Science Interview Question

Recommended companies

EXL › JPMC › Fractal Analytics › Mastercard › Innovaccer ›

EXL

Q: Differentiate between different built-in data types of Python?

There are four types of built-in data types in Python - list, tuple set and dictionary. List - They are Ordered, Changeable, and allow duplicates. They are declared using square brackets [], example - [‘Navin’, 6, 0.12]. Tuple - They are Ordered, Unchangeable, allow duplicates, they are indexed the same as list. They are declared using (), example - (‘Diwakar’, 18, 0.92). Set - They are Unordered, Unindexed, allow no duplicates, cannot be changed but we can add new items. They are declared using {}, example - {"apple", "banana", "cherry"}. Dictionary - They are Unordered, Changeable, indexed with no duplicates. Here items are stored in Key : Value pairs. Example - {"Rollno":32, "class":"X", "percentage":78.50}

Q: What is the difference between Correlation and Regression?

Correlation is to measure the degree of linear association between two variables. Here, the dependent and independent variables are assumed to be random. While in regression, we try to estimate or predict the average value of one variable on the basis of fixed values of other variables. Here, the dependent variable is assumed to be random while independent variables are assumed to be fixed in repeated sampling (independent variable can be intrinsically random)

Q: Explain Label Encoder and One Hot Encoder.

Label Encoder and One Hot Encoder are a part of the sci-kit learn library in Python, they are used to convert categorical data/text data into numbers, which our predictive model can better understand. Label Encoder converts categorical text data into model understandable numerical data. But, it introduces a new problem, where there is no relation in the categorical data but the model misunderstood the data to be in some kind of order. So to overcome this problem we use

one hot encoder. One Hot Encoder takes the label encoded categorical data column and then splits it into multiple columns which are called dummy variables.

Q: State the assumption of a Classical Linear Regression model.

The classical Linear Regression Model makes 7 assumptions: A1: The regression model is linear in the parameters, though it may or may not be linear in the variables. A2: The X variable(s) and the error term are independent, that is, cov(Xi,ui)=0. A3: The error term should follow a normal distribution; it means zero mean value and constant standard deviation. [It implies that there is no specification bias or specification error. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error.]. A4: Homoscedasticity: it means equal variance. There is no relationship between the error term and the predicted Y. A5: No Autocorrelation between the Disturbances: Given any two X values, Xi and Xj (i not = to j), the correlation between any two ui and uj is zero. A6: The Number of Observations n Must Be Greater than the Number of Parameters to Be Estimated (No. of explanatory variables). A7: No Multicollinearity: The independent variable should be linearly independent of each other.

Q: What is the difference between R-squared and Adjusted R-squared?

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. The problem with R square is that by adding more and more independent variables, irrespective of how well they are correlated to your dependent variable, R square increases. Obviously, this isn't a desirable property of a goodness-of-fit statistic R^2 = ESS/TSS = 1-RSS/TSS. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. But an independent variable that has a correlation to Y increases

adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. Adjusted R^2 will always be less than or equal to R^2. So, this is the desired property of a goodness-of-fit statistic. Adjusted R^2 = 1-((1-R^2)(n-1)/n-k-1)

Q: What is P-value?

When we do a regression analysis, we get a P-Value for each of the coefficients. Each term's p-value tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that the predictor changes are not associated with changes in the response. A high P-value is also called an insignificant P-value.

Q: Define cross-entropy loss function.

Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization in logistic regression. This function is known as the binary cross-entropy loss. Cost(h(x),y)= -log(h(x)) if y=1; -log(1-h(x)) if y=0. If you combine the above equations, we get a convex function that will help logistic regression to reach a global minimum faster. This cost function basically penalizes wrong predictions more than it diverts the right predictions.

Q: Explain Type 1 and Type 2 errors.

A type 1 error is also known as a false positive, the error of rejecting a null hypothesis when it is actually true. This means that you report that your findings are significant when in fact they have occurred by chance. The probability of making a type I error is called alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of

0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a type II error). A type II error is also known as a false negative, the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. Here a researcher concludes there is not a significant effect when actually there really is. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring your test has enough power.

Q: If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. The bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use a polling technique to combine all the predicted outcomes of the model. (You will read more about bagging algorithms in upcoming topics)

Q: What is the difference between overfitting and underfitting and how to identify them?

Overfitting: It happens when the model has focused too much on the training dataset that it cannot understand the test dataset. The Overfit model performs good (high accuracy) on training and bad (low accuracy) on tests.
Underfitting: It happens when the model has not captured the underlying logic of the data. Underfit model performs badly (low accuracy) on training and bad (low accuracy) on the test.

Q: What is the K-Means algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, firstly we select the value of K (the number of clusters) we want. Next, we select the random K points, the centroids (may or may not be from the datasets). Then, assign each data point to their closest centroid and calculate variance and place a new centroid of each cluster. Then reassign each data point to the new closest centroid, if any reassignment took place, we again calculate variance and place the new centroid of each cluster otherwise finish.

Q: What is a Neural Network?

Neural Networks are a type of machine learning algorithm which uses the concept of the human brain to facilitate the modeling of arbitrary functions. Neural Network requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. The most common Neural Networks consist of three network layers: an input layer, a hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and an output layer. Each sheet contains neurons called “nodes,” performing various operations. Neural Networks are used in deep learning algorithms like Convolutional NN, Recurrent NN, etc.

Contact Us

contact@thenoncore.com