ML Interview Questions

Common Interview Questions - Part 1

Linear Regression and General ML Questions

Question 1) What is linear regression?

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Question 2) State the assumptions in a linear regression model.

There are three main assumptions in a linear regression model:

1. **Assumption about the form of the model**:

   It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the '*linearity assumption*'.
2. **Assumptions about the residuals**:
    1. *Normality assumption*: It is assumed that the error terms, ε(i), are normally distributed.
    2. *Zero mean assumption*: It is assumed that the residuals have a mean value of zero.

3. *Constant variance assumption*: It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. *Independent error assumption*: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

3. **Assumptions about the estimators**:
    1. The independent variables are measured without error.
    2. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Explanation:

1. This is self-explanatory.
2. If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

   Also, the mean of the residuals should be zero.

   $$Y(i)i = \beta0 + \beta1x(i) + \varepsilon(i)$$

   This is the assumed linear model, where ε is the residual term.

   $$E(Y) = E(\beta0 + \beta1x(i) + \varepsilon(i))$$
   $$= E(\beta0 + \beta1x(i)) + E(\varepsilon(i))$$

   If the expectation(mean) of residuals, E(ε(i)), is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.

   The residuals (also known as error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.

3. If the independent variables are not linearly independent of each other, the uniqueness of the least squares solution (or normal equation solution) is lost.

Question 3) What is feature engineering? How do you apply it in the process of modelling?

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

In layman terms, feature engineering means the development of new features that may help you understand and model the problem in a better way. Feature engineering is essentially of two kinds — business driven and data-driven. Business-driven feature engineering revolves around the inclusion of features from a business point of view. The job here is to transform the business variables into features of the problem. In case of data-driven feature engineering, the features you add do not have any significant physical interpretation, but they help the model in the prediction of the target variable.

To apply feature engineering, one must be fully acquainted with the dataset. This involves knowing what the given data is, what it signifies, what the raw features are, etc. You must also have a crystal clear idea of the problem, such as what factors affect the target variable, what the physical interpretation of the variable is, etc.

Question 4) What is the use of regularisation? Explain L1 and L2 regularisations.

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. At times, the simple model

might not be able to generalise the data and the complex model overfits. To address this problem, regularisation is used.

Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalised and are small in magnitude. This essentially helps in capturing the trends in the data and at the same time prevents overfitting by not letting the model become too complex.

- **L1 or LASSO regularisation**: Here, the absolute values of the coefficients are added to the cost function. This can be seen in the following equation; the highlighted part corresponds to the L1 or LASSO regularisation. This regularisation technique gives sparse results, which lead to feature selection as well.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **L2 or Ridge regularisation**: Here, the squares of the coefficients are added to the cost function. This can be seen in the following equation, where the highlighted part corresponds to the L2 or Ridge regularisation.

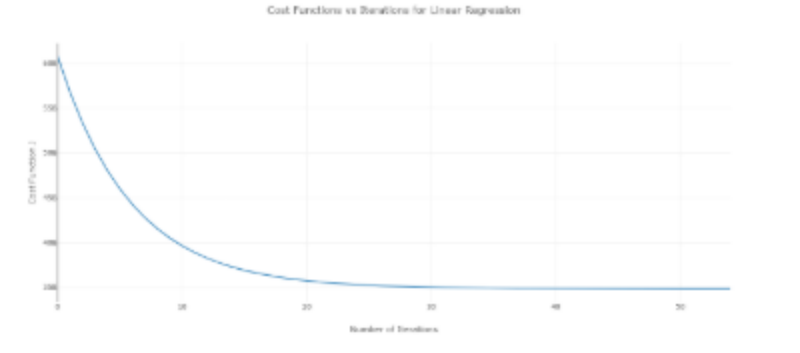
$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Question 5)How to choose the value of the parameter learning rate (α)?

Selecting the value of learning rate is a tricky business. If the value is too small, the gradient descent algorithm takes ages to converge to the optimal solution. On the other hand, if the value

of the learning rate is high, the gradient descent will overshoot the optimal solution and most likely never converge to the optimal solution.

To overcome this problem, you can try different values of alpha over a range of values and plot the cost vs the number of iterations. Then, based on the graphs, the value corresponding to the graph showing the rapid decrease can be chosen.



The aforementioned graph is an ideal cost vs number of iterations curve. Note that the cost initially decreases as the number of iterations increases, but after certain iterations, the gradient descent converges and the cost does not decrease anymore.

If you see that the cost is increasing with the number of iterations, your learning rate parameter is high and it needs to be decreased.

Common Interview Questions - Part 2

Question 1) How to choose the value of the regularisation parameter (λ)?

Selecting the regularisation parameter is a tricky business. If the value of λ is too high, it will lead to extremely small values of the regression coefficient, which will lead to the model underfitting (high bias - low variance). On the other hand, if the value of λ is 0 (very small), the model will tend to overfit the training data (low bias - high variance).

There is no proper way to select the value of λ. What you can do is have sub-samples of data and run the algorithm multiple times on different sets with values of λ. Here, the person has to decide how much variance can be tolerated. Once the user is satisfied with the variance, that value of λ can be chosen for the full dataset

One thing to be noted is that the value of λ selected here was optimal for that subset, not for the entire training data.

Question 2) Can we use linear regression for time series analysis?

One can use linear regression for time series analysis, but the results are not promising. So, it is generally not advisable to do so. The reasons behind this are —

1. Time series data is mostly used for the prediction of the future, but linear regression seldom gives good results for future prediction as it is not meant for extrapolation.

2. Mostly, time series data have a pattern, such as during peak hours, festive seasons, etc., which would most likely be treated as outliers in the linear regression analysis.

Question 3) What value is the sum of the residuals of a linear regression close to? Justify.

The sum of the residuals of a linear regression is 0. Linear regression works on the assumption that the errors (residuals) are normally distributed with a mean of 0, i.e.

$$Y=\beta_T X+\varepsilon$$

Here, Y is the target or dependent variable,

is the vector of the regression coefficient,

X is the feature matrix containing all the features as the columns,

ε is the residual term such that $\varepsilon \sim N(0,\sigma 2)$.

So, the sum of all the residuals is the expected value of the residuals times the total number of data points. Since the expectation of residuals is 0, the sum of all the residual terms is zero.

Note: $N(0,\sigma 2)$ is the standard notation for a normal distribution having mean μ and standard deviation σ2.

Question 4) How does multicollinearity affect the linear regression?

Multicollinearity occurs when some of the independent variables are highly correlated (positively or negatively) with each other. Multicollinearity causes a problem as it is against the

basic assumption of linear regression. The presence of multicollinearity does not affect the predictive capability of the model. So, if you just want predictions, the presence of multicollinearity does not affect your output. However, if you want to draw some insights from the model and apply them in, let's say, some business model, it may cause problems.

One of the major problems caused by multicollinearity is that it leads to incorrect interpretations and provides wrong insights. The coefficients of linear regression suggest the mean change in the target value if a feature is changed by one unit. So, if multicollinearity exists, this does not hold true as changing one feature will lead to changes in the correlated variable and consequent changes in the target variable. This leads to wrong insights and can produce hazardous results for a business.

A highly effective way of dealing with multicollinearity is the use of VIF (Variance Inflation Factor). Higher the value of VIF for a feature, more linearly correlated is that feature. Simply remove the feature with very high VIF value and re-train the model on the remaining dataset.

Question 5) What is the normal form (equation) of linear regression? When should it be preferred to the gradient descent method?

The normal equation for linear regression is —

$$\beta=(X^TX)^{-1}X^TY$$

Here, $Y=\beta^TX$ is the model for the linear regression,

Y is the target or dependent variable,

β is the vector of the regression coefficient, which is arrived at using the normal equation,

X is the feature matrix containing all the features as the columns.

Note here that the first column in the X matrix consists of all 1s. This is to incorporate the offset value for the regression line.

Comparison between gradient descent and normal equation:

Gradient Descent	**Normal Equation**
Needs hyper-parameter tuning for alpha (learning parameter)	No such need
It is an iterative process	It is a non-iterative process
$O(kn^2)$ time complexity	O(n3) time complexity due to the evaluation of X^TX
Prefered when n is extremely large	Becomes quite slow for large values of n

Here, ‘k’ is the maximum number of iterations for gradient descent, and ‘n’ is the total number of data points in the training set.

Clearly, if we have large training data, normal equation is not prefered for use. For small values of ‘n’, normal equation is faster than gradient descent.

Question 6) You run your regression on different subsets of your data, and in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

This case implies that the dataset is heterogeneous. So, to overcome this problem, the dataset should be clustered into different subsets, and then separate models should be built for each cluster. Another way to deal with this problem is to use non-parametric models, such as decision trees, which can deal with heterogeneous data quite efficiently.

Question 7) Your linear regression doesn't run and communicates that there is an infinite number of best estimates for the regression coefficients. What could be wrong?

This condition arises when there is a perfect correlation (positive or negative) between some variables. In this case, there is no unique value for the coefficients, and hence, the given condition arises.

Question 8) What do you mean by adjusted $R2$? How is it different from $R2$?

Adjusted $R2$, just like $R2$, is a representative of the number of points lying around the regression line. That is, it shows how well the model is fitting the training data. The formula for adjusted $R2$ is

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

Here, n is the number of data points, and k is the number of features.

One drawback of R^2 is that it will always increase with the addition of a new feature, whether the new feature is useful or not. The adjusted R^2 overcomes this drawback. The value of the adjusted R^2increases only if the newly added feature plays a significant role in the model.

Common Interview Questions - Part 3

Question 1) How do you interpret the residual vs fitted value curve?

The residual vs fitted value plot is used to see whether the predicted values and residuals have a correlation or not. If the residuals are distributed normally, with a mean around the fitted value and a constant variance, our model is working fine; otherwise, there is some issue with the model.

The most common problem that can be found when training the model over a large range of a dataset is heteroscedasticity (this is explained in the answer below). The presence of heteroscedasticity can be easily seen by plotting the residual vs fitted value curve.

Question 2) What is heteroscedasticity? What are the consequences, and how can you overcome it?

A random variable is said to be heteroscedastic when different sub-populations have different variabilities (standard deviation).

The existence of heteroscedasticity gives rise to certain problems in the regression analysis as the assumption says that error terms are uncorrelated and, hence, the variance is constant. The presence of heteroscedasticity can often be seen in the form of a cone-like scatter plot for residual vs fitted values.

One of the basic assumptions of linear regression is that heteroscedasticity is not present in the data. Due to violation of the assumptions, the Ordinary Least Squares (OLS) estimators are not the Best Linear Unbiased Estimators (BLUE). Hence, they do not give the least variance than other Linear Unbiased Estimators (LUEs).

There is no fixed procedure to overcome heteroscedasticity. However, there are some ways that may lead to the reduction of heteroscedasticity. They are —

1. *Logarithmising the data*: A series that is increasing exponentially often results in increased variability. This can be overcome using the log transformation.
2. *Using weighted linear regression*: Here, the OLS method is applied to the weighted values of X and Y. One way is to attach weights directly related to the magnitude of the dependent variable.

Question 3) What is VIF? How do you calculate it?

Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset. It is calculated as —

$$VIF_j = \frac{1}{1 - R2_j}$$

Here, VIF_j is the value of VIF for the j'th variable,

$R2_j$ is the $R2$ value of the model when that variable is regressed against all the other independent variables.

If the value of VIF is high for a variable, it implies that the $R2$ value of the corresponding model is high, i.e. other independent variables are able to explain that variable. In simple terms, the variable is linearly dependent on some other variables.

Question 4) How do you know that linear regression is suitable for any given data?

To see if linear regression is suitable for any given data, a scatter plot can be used. If the relationship looks linear, we can go for a linear model. But if it is not the case, we have to apply some transformations to make the relationship linear. Plotting the scatter plots is easy in case of simple or univariate linear regression. But in case of multivariate linear regression, two dimensional pair-wise scatter plots, rotating plots, and dynamic graphs can be plotted.

Question 5) How is hypothesis testing used in linear regression?

Hypothesis testing can be carried out in linear regression for the following purposes:

1. To check whether a predictor is significant for the prediction of the target variable. Two common methods for this are —
    - *By the use of p-values*:

      If the p-value of a variable is greater than a certain limit (usually 0.05), the variable is insignificant in the prediction of the target variable.

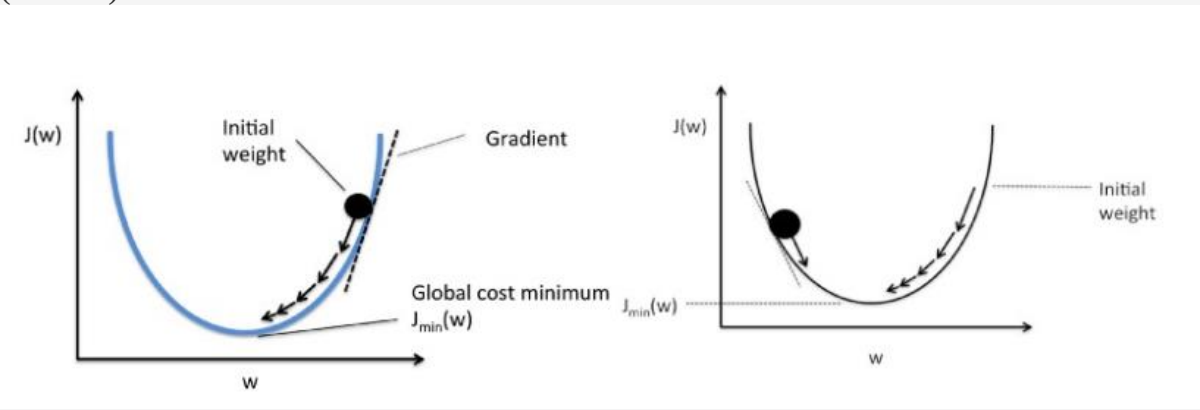
    - *By checking the values of the regression coefficient*:

      If the value of regression coefficient corresponding to a predictor is zero, that variable is insignificant in the prediction of the target variable and has no linear relationship with it.
2. To check whether the calculated regression coefficients are good estimators of the actual coefficients.

Question 6) Explain gradient descent with respect to linear regression.

Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin $J(\theta_0,\theta_1)$, where $J(\theta_0,\theta_1)$ is the cost function of the linear regression. It is given by

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Here, h is the linear hypothesis model, $h=\theta_0+\theta_1 X$,

y is the true output,

and m is the number of the data points in the training set.

Common Interview Questions - Part 4

Question 1) How do you interpret a linear regression model?

A linear regression model is quite easy to interpret. The model is of the following form:

$$y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n$$

The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of xi increases by 1 unit, keeping other variables constant, the total increase in the value of y will be β_i . Mathematically, the intercept term (β_0) is the response when all the predictor terms are set to zero or not considered.

Question 2) What is robust regression?

A regression model should be robust in nature. This means that with changes in a few observations, the model should not change drastically. Also, it should not be much affected by the outliers.

A regression model with OLS (Ordinary Least Squares) is quite sensitive to the outliers. To overcome this problem, we can use the WLS (Weighted Least Squares) method to determine the estimators of the regression coefficients. Here, less weights are given to the outliers or high leverage points in the fitting, making these points less impactful.

Question 3) Which graphs are suggested to be observed before model fitting?

Before fitting the model, one must be well aware of the data, such as what the trends, distribution, skewness, etc. in the variables are. Graphs such as histograms, box plots, and dot plots can be used to observe the distribution of the variables. Apart from this, one must also analyse what the relationship between dependent and independent variables is. This can be done by scatter plots (in case of univariate problems), rotating plots, dynamic plots, etc.

Question 5) What is the generalized linear model?

The generalized linear model is the derivative of the ordinary linear regression model. GLM is more flexible in terms of residuals and can be used where linear regression does not seem appropriate. GLM allows the distribution of residuals to be other than a normal distribution. It generalizes the linear regression by allowing the linear model to link to the target variable using the linking function. Model estimation is done using the method of maximum likelihood estimation.

Question 6) Explain the bias-variance trade-off.

Bias refers to the difference between the values predicted by the model and the real values. It is an error. One of the goals of an ML algorithm is to have low bias.

Variance refers to the sensitivity of the model to small fluctuations in the training dataset. Another goal of an ML algorithm is to have low variance.

For a dataset that is not exactly linear, it is not possible to have both bias and variance low at the same time. A straight line model will have low variance but high bias, whereas a high-degree polynomial will have low bias but high variance.

There is no escaping the relationship between bias and variance in machine learning.

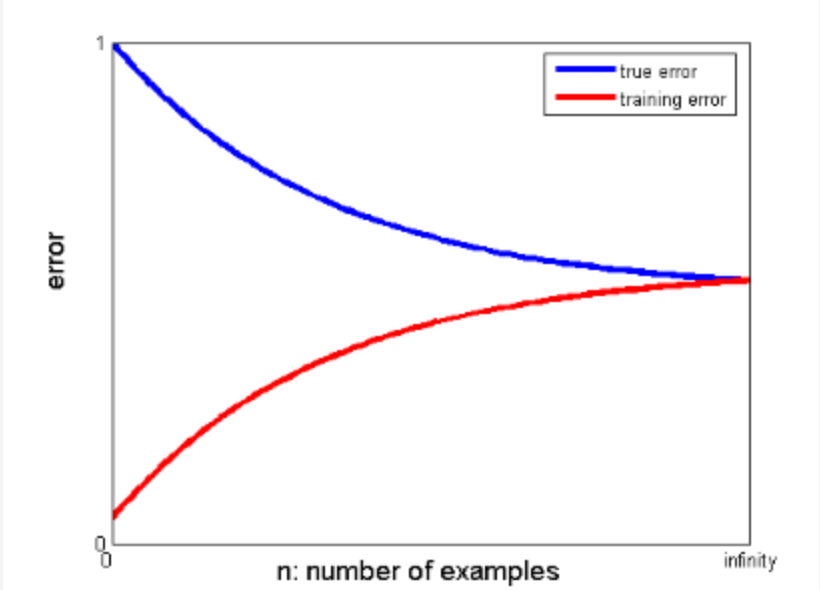
1. Decreasing the bias increases the variance.
2. Decreasing the variance increases the bias.

So, there is a trade-off between the two; the ML specialist has to decide, based on the assigned problem, how much bias and variance can be tolerated. Based on this, the final model is built.

Question 7) How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.

In a learning curve, the training error and cross-validating error are plotted against the number of training data points. A typical learning curve looks like this:



If the training error and true error (cross-validating error) converge to the same value and the corresponding value of error is high, it indicates that the model is underfitting and is suffering from high bias.

If there is a significant gap between the converging values of the training and cross-validating errors, i.e. the cross-validating error is significantly higher than the training error, it suggests that the model is overfitting the training data and is suffering from high variance.

Question 8) Your model is not working well. Is getting more data always a solution?

No, getting more data is not always a solution. If the model is suffering from high bias, getting more data will not help after a point of time. Also, the more the training data, the more is the storage required. This leads to more computational power being needed for training the model. So, before trying to get more data, the cost associated with it must also be considered.

Question 9) What are parametric and non-parametric machine learning algorithms?

Assumptions can greatly simplify the learning process but can also limit what can be learned. Algorithms that simplify a function to a known form are called parametric machine learning algorithms. Algorithms that do not make strong assumptions about the form of a mapping function are called non-parametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data.

Examples of parametric machine learning algorithms are —

- Logistic regression
- Linear discriminant analysis
- Naive Bayes
- Simple neural networks

Examples of non-parametric machine learning algorithms are —

- K-Nearest Neighbors
- Decision trees, such as CART and C4.5
- Support vector machines

Logistic Regression

Common Interview Questions - Part 1

Question 1) What is a logistic function? What is the range of values of a logistic function? $f(z)=\frac{1}{1+e^{-z}}$

The values of a logistic function will range from 0 to 1. The values of Z will vary from -infinity to +infinity.

Question 2) Why is logistic regression very popular?

Logistic regression is famous because it can convert the values of logits (log-odds), which can range from -infinity to +infinity to a range between 0 and 1. As logistic functions output the probability of occurrence of an event, it can be applied to many real-life scenarios. It is for this reason that the logistic regression model is very popular.

Question 3) What is the formula for the logistic regression function?

$$f(z)=\frac{1}{1+e^{\alpha+\beta 1X1+\beta 2X2+....+\beta kXk}}$$

Question 4) How can the probability of a logistic regression model be expressed as conditional probability?

P(Discrete value of Target variable | X1, X2, X3….Xk). It is the probability of the target variable to take up a discrete value (either 0 or 1 in case of binary classification problems) when the values of independent variables are given. For example, the probability an employee will attrite (target variable) given his attributes such as his age, salary, KRA's, etc.

Question 5) What are odds?

It is the ratio of the probability of an event occurring to the probability of the event not occurring. For example, let's assume that the probability of winning a lottery is 0.01. Then, the probability of not winning is 1- 0.01 = 0.99.

The odds of winning the lottery = probability of winningprobability of not winning

The odds of winning the lottery = 0.010.99

The odds of winning the lottery is 1 to 99, and the odds of not winning the lottery is 99 to 1.

Question 6) What are the outputs of the logistic model and the logistic function?

The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities.

Logistic model = $\alpha+\beta_1X_1+\beta_2X_2+....+\beta_kX_k$ The output of the same will be logits.

Logistic function = $f(z)=\frac{1}{1+e^{\alpha+\beta_1X_1+\beta_2X_2+....+\beta_kX_k}}$. The output, in this case, will be the probabilities.

Question 7) How to interpret the results of a logistic regression model? Or, what are the meanings of alpha and beta in a logistic regression model?

Alpha is the baseline in a logistic regression model. It is the log odds for an instance when all the attributes (X1, X2,.....Xk) are zero. In practical scenarios, the probability of all the attributes being zero is very low. In another interpretation, Alpha is the log odds for an instance when none of the attributes is taken into consideration.

Beta is the value by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

Question 8) What is odds ratio?

Odds ratio is the ratio of odds between two groups. For example, let's assume that we are trying to ascertain the effectiveness of a medicine. We administered this medicine to the 'intervention' group and a placebo to the 'control' group.

Odds ratio (OR) = odds of the intervention group/odds of the control group

Interpretation

If odds ratio = 1, then there is no difference between the intervention group and the control group

If odds ratio is greater than 1, then the control group is better than the intervention group

If odds ratio is less than 1, then the intervention group is better than the control group.

Question 9) What is the formula for calculating odds ratio?

$$\mathrm{OR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

In the formula above, X1 and X0 stand for two different groups for which odds ratio needs to be calculated. X1i stands for the instance 'i' in group X1. X0i stands for the instance 'i' in group X0. βi stands for the coefficient of the logistic regression model. Note that the baseline is not included in this formula.

Question 10) Why can't linear regression be used in place of logistic regression for binary classification?

The reasons why linear regressions cannot be used in case of binary classification are as follows:

- *Distribution of error terms*: The distribution of data in case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In case of binary classification, this assumption does not hold true.
- *Model output*: In linear regression, the output is continuous. In case of binary classification, an output of a continuous value does not make sense. For binary classification problems, linear regression may predict values that can go beyond 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different

classes, then its range should be restricted to 0 and 1. As the logistic regression model can output probabilities with logistic/sigmoid function, it is preferred over linear regression.

- *Variance of Residual errors*: Linear regression assumes that the variance of random errors is constant. This assumption is also violated in case of logistic regression.

Common Interview Questions - Part 2

Question 1) 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.

Logistic regression model formula = α+1X1+2X2+....+kXk. This clearly represents a straight line. Logistic regression is only suitable in such cases where a straight line is able to separate the different classes. If a straight line is not able to do it, then nonlinear algorithms should be used to achieve better results.

Question 2) What is the likelihood function?

The likelihood function is the joint probability of observing the data. For example, let's assume that a coin is tossed 100 times and we want to know the probability of getting 60 heads from the tosses. This example follows the binomial distribution formula.

p = Probability of heads from a single coin toss

n = 100 (the number of coin tosses)

x = 60 (the number of heads - success)

n-x = 40 (the number of tails)

Pr(X=60 |n = 100, p)

The likelihood function is the probability that the number of heads received is 60 in a trail of 100 coin tosses, where the probability of heads received in each coin toss is p. Here the coin tosses follow the binomial distribution.

This can be reframed as follows:

Pr(X=60|n=100,p) = c x p60x(1-p)100-60 = L()

c = constant

p = unknown parameter

The likelihood function gives the probability of observing the results using unknown parameters.

Question 3) What is the Maximum Likelihood Estimator (MLE)?

The MLE chooses those sets of unknown parameters (estimator) that maximise the likelihood function. The method to find the MLE is to use calculus and setting the derivative of the logistic

function with respect to an unknown parameter to zero, and solving it will give the MLE. For a binomial model, this will be easy, but for a logistic model, the calculations are complex. Computer programs are used for deriving MLE for logistic models.

(Here's another approach to answering the question.)

MLE is a statistical approach to estimating the parameters of a mathematical model. MLE and ordinary square estimation give the same results for linear regression if the dependent variable is assumed to be normally distributed. MLE does not assume anything about independent variables.

Question 4) What are the different methods of MLE and when is each method preferred?

In case of logistics regression, there are two approaches of MLE. They are conditional and unconditional methods. Conditional and unconditional methods are algorithms that use different likelihood functions. The unconditional formula employs joint probability of positives (for example, churn) and negatives (for example, non-churn). The conditional formula is the ratio of the probability of observed data to the probability of all possible configurations.

The unconditional method is preferred if the number of parameters is lower compared to the number of instances. If the number of parameters is high compared to the number of instances, then conditional MLE is to be preferred. Statisticians suggest that conditional MLE is to be used when in doubt. Conditional MLE will always provide unbiased results.

Question 5) What are the advantages and disadvantages of conditional and unconditional methods of MLE?

Conditional methods do not estimate unwanted parameters. Unconditional methods estimate the values of unwanted parameters also. Unconditional formulas can directly be developed with joint probabilities. This cannot be done with conditional probability. If the number of parameters is high relative to the number of instances, then the unconditional method will give biased results. Conditional results will be unbiased in such cases.

Question 6) What is the output of a standard MLE program?

The output of a standard MLE program is as follows:

- **Maximised likelihood value**: This is the numerical value obtained by replacing the unknown parameter values in the likelihood function with the MLE parameter estimator.
- **Estimated variance-covariance matrix**: The diagonal of this matrix consists of estimated variances of the ML estimates. The off-diagonal consists of the covariances of the pairs of the ML estimates.

Question 7) Why can't we use Mean Square Error (MSE) as a cost function for logistic regression?

In logistic regression, we use the sigmoid function and perform a non-linear transformation to obtain the probabilities. Squaring this non-linear transformation will lead to non-convexity with local minimums. Finding the global minimum in such cases using gradient descent is not possible. Due to this reason, MSE is not suitable for logistic regression. Cross-entropy or log loss

is used as a cost function for logistic regression. In the cost function for logistic regression, the confident wrong predictions are penalised heavily. The confident right predictions are rewarded less. By optimising this cost function, convergence is achieved.

Common Interview Questions - Part 3

Question 1) Why is accuracy not a good measure for classification problems?

Accuracy is not a good measure for classification problems because it gives equal importance to both false positives and false negatives. However, this may not be the case in most business problems. For example, in case of cancer prediction, declaring a cancer as benign is more serious than wrongly informing the patient that he is suffering from cancer. Accuracy gives equal importance to both cases and cannot differentiate between them.

Question 2) What is the importance of a baseline in a classification problem?

Most classification problems deal with imbalanced datasets. Examples include telecom churn, employee attrition, cancer prediction, fraud detection, online advertisement targeting, and so on. In all these problems, the number of the positive classes will be very low when compared to the negative classes. In some cases, it is common to have positive classes that are less than 1% of the total sample. In such cases, an accuracy of 99% may sound very good but, in reality, it may not be. Here, the negatives are 99%, and hence, the baseline will remain the same. If the algorithms predict all the instances as negative, then also the accuracy will be 99%. In this case, all the positives will be predicted wrongly, which is very important for any business. Even though all

the positives are predicted wrongly, an accuracy of 99% is achieved. So, the baseline is very important, and the algorithm needs to be evaluated relative to the baseline.

Question 3) What are false positives and false negatives?

False positives are those cases in which the negatives are wrongly predicted as positives. For example, predicting that a customer will churn when, in fact, he is not churning.

False negatives are those cases in which the positives are wrongly predicted as negatives. For example, predicting that a customer will not churn when, in fact, he churns.

Question 4) What are the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR)?

TPR refers to the ratio of positives correctly predicted from all the true labels. In simple words, it is the frequency of correctly predicted true labels.

$$TPR = TP/TP+FN$$

TNR refers to the ratio of negatives correctly predicted from all the false labels. It is the frequency of correctly predicted false labels.

$$TNR = TN/TN+FP$$

FPR refers to the ratio of positives incorrectly predicted from all the true labels. It is the frequency of incorrectly predicted false labels.

$$FPR = FP/TN+FP$$

FNR refers to the ratio of negatives incorrectly predicted from all the false labels. It is the frequency of incorrectly predicted true labels.

$$FNR = FN/TP+FN$$

Question 5) What are precision and recall?

Precision is the proportion of true positives out of predicted positives. To put it in another way, it is the accuracy of the prediction. It is also known as the 'positive predictive value'.

$$Precision = TP/TP+FP$$

Recall is same as the true positive rate (TPR).

Question 6) What is F-measure?

It is the harmonic mean of precision and recall. In some cases, there will be a trade-off between the precision and the recall. In such cases, the F-measure will drop. It will be high when both the precision and the recall are high. Depending on the business case at hand and the goal of data analytics, an appropriate metric should be selected.

$$F\text{-measure} = 2* (Precision * Recall \ / Precision+Recall)$$

Question 7) What is accuracy?

It is the number of correct predictions out of all predictions made.

Accuracy = TP+TN / The total number of Predictions

Question 8) What are sensitivity and specificity?

Sensitivity is the true positive rate.

Sensitivity = TP / TP + FN

Specificity is the same as true negative rate, or it is equal to 1 - false positive rate.

Specificity = TN / TN + FP.

Question 9) How to choose a cutoff point in case of a logistic regression model?

The cutoff point depends on the business objective. Depending on the goals of your business, the cutoff point needs to be selected. For example, let's consider loan defaults. If the business objective is to reduce the loss, then the specificity needs to be high. If the aim is to increase the profits, then it is an entirely different matter. It may not be the case that profits will increase by avoiding giving loans to all predicted default cases. But it may be the case that the business has to disburse loans to default cases that are slightly less risky to increase the profits. In such a case, a different cutoff point, which maximises profit, will be required. In most of the instances, businesses will operate around many constraints. The cutoff point that satisfies the business objective will not be the same with and without limitations. The cutoff point needs to be selected considering all these points. As a thumb rule, choose a cutoff value that is equivalent to the proportion of positives in a dataset.

Question 10) How does logistic regression handle categorical variables?

The inputs to a logistic regression model need to be numeric. The algorithm cannot handle categorical variables directly. So, they need to be converted into a format that is suitable for the algorithm to process. The various levels of a categorical variable will be assigned a unique numeric value known as the dummy variable. These dummy variables are handled by the logistic regression model as any other numeric value.

Common Interview Questions - Part 4

Question 1) What are lift curves?

The lift is the improvement in model performance (increase in true positive rate) when compared to random performance. Random performance means if 50% of the instances is targeted, then it is expected that it will detect 50% of the positives. Lift is in comparison to the random performance of a model. If a model's performance is better than its random performance, then its lift will be greater than 1.

In a lift curve, lift is plotted on the Y-axis and the percentage of the population (sorted in descending order) on the X-axis. At a given percentage of the target population, a model with a high lift is preferred.

Question 2) Which algorithm is better at handling outliers logistic regression or SVM?

Logistic regression will find a linear boundary if it exists to accommodate the outliers. Logistic regression will shift the linear boundary in order to accommodate the outliers. SVM is

insensitive to individual samples. There will not be a major shift in the linear boundary to accommodate an outlier. SVM comes with inbuilt complexity controls, which take care of overfitting. This is not true in case of logistic regression.

Question 3) How will you deal with the multiclass classification problem using logistic regression?

The most famous method of dealing with multiclass classification using logistic regression is using the one-vs-all approach. Under this approach, a number of models are trained, which is equal to the number of classes. The models work in a specific way. For example, the first model classifies the datapoint depending on whether it belongs to class 1 or some other class; the second model classifies the datapoint into class 2 or some other class. This way, each data point can be checked over all the classes.

Question 4) Explain the use of ROC curves and the AUC of an ROC Curve.

An ROC (Receiver Operating Characteristic) curve illustrates the performance of a binary classification model. It is basically a TPR versus FPR (true positive rate versus false positive rate) curve for all the threshold values ranging from 0 to 1. In an ROC curve, each point in the ROC space will be associated with a different confusion matrix. A diagonal line from the bottom-left to the top-right on the ROC graph represents random guessing. The Area Under the Curve (AUC) signifies how good the classifier model is. If the value for AUC is high (near 1), then the model is working satisfactorily, whereas if the value is low (around 0.5), then the model is not working properly and just guessing randomly.

Question 5) How can you use the concept of ROC in a multiclass classification?

The concept of ROC curves can easily be used for multiclass classification by using the one-vs-all approach. For example, let's say that we have three classes ‘a’, ’b’, and ‘c’. Then, the first class comprises class ‘a’ (true class) and the second class comprises both class ‘b’ and class ‘c’ together (false class). Thus, the ROC curve is plotted. Similarly, for all the three classes, we will plot three ROC curves and perform our analysis of AUC.

Question 6) What is a cumulative response curve (CRV)?

In order to convey the results of an analysis to management, a ‘cumulative response curve’ is used, which is more intuitive than the ROC curve. An ROC curve is very difficult to understand for someone outside the field of data science. A CRV consists of the true positive rate or the percentage of positives correctly classified on the Y-axis and the percentage of the population targeted on the X-axis. It is important to note that the percentage of the population will be ranked by the model in descending order (either the probabilities or the expected values). If the model is good, then by targeting a top portion of the ranked list, all high percentages of positives will be captured. As with the ROC curve, there will be a diagonal line which represents random performance. Let's understand this random performance with an example. Assuming that 50% of the list is targeted, it is expected that it will capture 50% of the positives. This expectation is captured by the diagonal line, which is similar to the ROC curve.

Clustering

Common Interview Questions - Part 1

Question 1) What is clustering and why is it used?

Clustering is an unsupervised technique used to group data objects. The groups are created in such a way that the points in a group are more similar to each other and different from the points in another group. The purpose of clustering can be divided into two broad categories, understanding and utility.

Clustering is done to understand the natural clusters or patterns that exist in data. For example, a business that uses clustering to segment its customer base needs to understand the characteristics of different segments.

Clustering is also done for utility. In this case, a prototype is created to act as a representation of the data objects in a particular cluster. This prototype is used for the purpose of analysis. For example, summarisation. If the dataset is very large, then it becomes very difficult to apply machine learning algorithms to it. In such cases, clustering can be used to create prototypes. Machine learning algorithms are used on the reduced dataset, which only consists of cluster prototypes. Other examples of using prototypes for utility include compressing images, finding nearest neighbours efficiently, etc.

Question 2) What are the different types of clustering?

Hierarchical versus partitional: Hierarchical clustering forms nested clusters. In this type, the subclusters exist in other clusters. In the partitional type, subclusters do not exist, and all the data objects belong to non-overlapping clusters.

Exclusive, overlapping, versus fuzzy: In the exclusive type cluster, a data object will exclusively belong to a single cluster. In case of an overlapping cluster, a data object will belong to multiple clusters. For example, an UpGrad employee who is also a student of one of UpGrad's programs will belong to both the 'employee' and 'student' categories. In a fuzzy cluster, a data object will belong to all the clusters with a probability. The probability of a point belonging to different clusters will be different, and it will sum up to 1.

Complete versus partial: Complete is a type of cluster in which all the data objects are assigned to a cluster. In the partial type, data objects are left out without getting assigned to any cluster. For example, leaving out outliers without assigning them to any cluster.

Question 3) What are the different types of clusters?

Well-separated: These are clusters in which data objects exclusively belong to a single cluster.

Prototype: The data objects in this type of cluster are more similar to the prototype of the cluster it belongs to than to the prototype of a different cluster.

Graph: In this type, the data is represented in the form of a graph. The nodes will be the data objects, and the edges will represent the connections between the data objects.

Density: Denser regions are clustered separately from the less dense regions of data objects.

Conceptual: The data points in a cluster of this type share the same property. In this type of clustering, the algorithm searches for a specific concept to group the data objects together.

Question 4) What is K-means clustering? How does its algorithm work?

It is a prototype-based partition algorithm. The K-means algorithm works by finding out a centroid, which acts as a prototype for its cluster. The centroid for the K-means algorithm will be the mean of all the data points in a particular cluster. As the centroid will be the mean, it may not correspond to any data point in the cluster.

The algorithm for K-means algorithm is as follows:

- Select initial centroids. The input regarding the number of centroids should be given by the user.
- Assign the data points to the closest centroid
- Recalculate the centroid for each cluster and assign the data objects again
- Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of clusters.

Question 5) What are the different proximity functions or distance metrics used for the K-means algorithm?

Euclidean, Manhattan, Cosine, and Bregman divergence are some distance metrics used for the K-means algorithm. Euclidean is the squared distance from a data point to the centroid. Manhattan is the absolute distance from a data point to the centroid. Cosine is the cosine distance from a data point to the cluster centroid. Bregman divergence is a class of distance metrics that includes Euclidean, Mahalanobis, and Cosine. Basically, Bregman divergence includes all those distance metrics for which the mean is a centroid.

Common Interview Questions - Part 2

Question 1) What are the issues with random initialisation of centroids in K-means algorithm and how to overcome it?

Initiation of the centroids in a cluster is one of the most important steps of the K-means algorithm. Many times, random selection of initial centroid does not lead to an optimal solution. In order to overcome this problem, the algorithm is run multiple times with different random initialisations. The sum of squared errors (SSE) are calculated for different initial centroids. The set of centroids with the minimum SSE is selected. Even though this is a very simple method, it is not foolproof. The results of multiple random cluster initialisations will depend on the dataset and the number of clusters selected, however, that still will not give an optimum output every time.

The other method involves first selecting the centroid of all the data points. Then, for each successive initial centroid, select the point which is the farthest from the already selected centroid. This procedure will ensure that the selection is random, and the centroids are far apart. The disadvantage of this method is that calculating the farthest point will be expensive. In order to avoid this problem, initialisation is carried out on a subset of the dataset.

The other methods of handling this are bisecting K-means (this is covered in another question below) and taking care of the issues once clustering is done post processing.

Question 2) How are outliers handled by the K-means algorithm?

Handling of outliers differs from case to case. In some cases, it will provide very useful information, and in some cases, it will severely affect the results of the analysis. Having said

that, let’s learn about some of the issues that arise due to outliers in the K-means algorithm below.

The centroids will not be a true representation of a cluster in the presence of outliers. The sum of squared errors (SSE) will also be very high in case of outliers. Small clusters will bond with outliers, which may not be the true representation of the natural patterns of clusters in data. Due to these reasons, outliers need to be removed before proceeding with clustering on the data.

Question 3) What is the objective function for measuring the quality of clustering in case of the K-means algorithm with Euclidean distance?

Sum of squared errors (SSE) is used as the objective function for K-means clustering with Euclidean distance. The Euclidean distance is calculated from each data point to its nearest centroid. These distances are squared and summed to obtain the SSE. The aim of the algorithm is to minimise the SSE. Note that SSE considers all the clusters formed using the K-means algorithm.

Question 4) What is Bisecting K-means?

In Bisecting K-means, all the data objects are split into two clusters. One of these clusters is selected and divided into two groups. This will continue until a pre-decided K number of clusters is formed. The selection of the cluster for splitting will depend on a number of factors such as the size of the cluster, the largeness of the SSE, etc. The choice of the splitting criteria will decide the type of clusters formed. The centroids of Bisecting K-means is used by the K-means algorithm as initial centroids for further refining the clusters. K-means algorithm will be able to

find the local minimum by using the SSE as the objective function. In case of Bisecting K-means, the local optimum is related to the clusters being split and not the complete dataset. So, the final results obtained will not be ‘local’ to the dataset with respect to the total SSE. Bisecting K-means does not have issues with initialisation because when different splits are tried out, a split with a low SSE is preferred. Additionally, Bisecting K-means only deals with two centroids at a time. Using Bisecting K-means as a base for K-means improves the performance of the latter.

Question 5) Is K-means clustering suitable for all shapes and sizes of clusters?

K-means is not suitable for all shapes, sizes, and densities of clusters. If the natural clusters of a dataset are vastly different from a spherical shape, then K-means will face great difficulties in detecting it. K-means will also fail if the sizes and densities of the clusters are different by a large margin. This is mostly due to using SSE as the objective function, which is more suited for spherical shapes. SSE is not suited for clusters with non-spherical shapes, varied cluster sizes, and densities.

Common Interview Questions - Part 3

Question 1) What are the advantages and disadvantages of K-means clustering?

K-means is a simple and efficient algorithm though it requires multiple runs. Bisecting K-means, a flavour of K-means, is more efficient and handles initialisation well. The disadvantages of K-means is that they can handle all shapes and sizes of clusters, but it is not good at handling outliers. This algorithm is only applicable to the data for which a centroid exists.

Question 2) What is hierarchical clustering?

There are two types of hierarchical clustering. They are agglomerative clustering and divisive clustering.

Agglomerative clustering: In this algorithm, initially every data object will be treated as a cluster. In each step, the nearest clusters will fuse together and form a bigger cluster. Ultimately, all the clusters will merge together. Finally, a single cluster, which encompasses all the data points, will remain.

Divisive clustering: This is the opposite of the agglomerative clustering. In this type, all the data objects will be considered as single clusters. In each step, the algorithm will split the cluster. This will repeat until only single data points remain, which will be considered as singleton clusters.

Question 3) What are the proximity measures between clusters and when is it to be used?

Single link, complete link, group average, and ward's are some of the proximity measures for hierarchical clustering.

Single link: Cluster proximity in case of single linkage is the distance of two nearest points in two different clusters. This method is good at handling non-elliptical shapes. However, it is susceptible to noise and outliers.

Complete link: It is the distance between the two farthest points of two different clusters. It favours globular shapes and is not susceptible to noise and outliers. This method of proximity measurement breaks large clusters.

Group average: This cluster proximity is the average of all pairwise distances (pairwise distance is the distance between two points in two different clusters; all combinations of pairs are considered for calculating the average). This is an intermediate approach between the single and complete linkages.

Ward's method: The proximity between two clusters is the increase in SSE that is the result of merging both the clusters.

Question 4) What are the disadvantages of agglomerative hierarchical clustering?

Objective function: SSE is the objective function for K-means. Likewise, there exists no global objective function for hierarchical clustering. It considers proximity locally before merging two clusters.

Time and space complexity: The time and space complexity of agglomerative clustering is more than K-means clustering, and in some cases, it is prohibitive.

Final merging decisions: The merging decisions, once given by the algorithm, cannot be undone at a later point in time. Due to this, a local optimisation criteria cannot become global criteria. Note that there are some advanced approaches available to overcome this problem.

Question 5) What are the strengths and weaknesses of the agglomerative hierarchical algorithm?

This algorithm is helpful where a hierarchy is required for resolving a business problem. Research suggests that this algorithm gives better results. But it has some disadvantages such as high storage and time costs. For this particular algorithm, as merges once created are final, it is not suitable for high-dimensional and noisy data.

Question 6) Is validation required for clustering? If yes, then why is it required?

Clustering algorithms have a tendency to cluster even when the data is random. It is essential to validate if a non-random structure is present in the data. It is also required to validate whether the number of clusters formed is appropriate or not. Evaluation of clusters is done with or without external reference to check the fitness of the data. Evaluation is also done to compare clusters and decide the better among them.

Question 7) What are the different types of validation or evaluation measures?

Two different types of cluster validation are unsupervised and supervised methods.

Unsupervised: In the unsupervised method, there is no reference to the external information. Due to this, unsupervised methods are also known as internal indices. Unsupervised methods are further divided into two methods, cluster cohesion and cluster separation. Cluster cohesion is the measure of tightness between the elements of a cluster. Cluster separation is the measure of how well separated one cluster is from the other clusters.

Supervised: In supervised methods, the cluster is evaluated with external information provided. Due to this, they are known as external indices. Entropy is one such measure. It compares how the cluster labels fare with externally supplied label information.

Relative measures are those which make use of either supervised or unsupervised methods to compare different clusters. For example, two K-mean clusters compared with the help of SSE.

Question 8) How are prototype clusters evaluated using cohesion and separation?

In prototype clustering, cohesion is defined as the sum of proximities between the data objects and the centroid. Separation is of two kinds. It is either the proximity between the two cluster centroids or the proximity between the cluster centroid and the overall centroid of all the data objects.

Cohesion = Sum of proximities between data objects and the centroid

Separation = proximity between the two cluster centroids (or)

Separation = proximity between the cluster centroid and the overall centroid.

An overall validity of cluster validation is calculated with a weighted sum of cohesion and separation measures. Different kinds of weights can be used for calculating the overall measure. Typically, these weights are a measure of cluster sizes.

Question 9) How to decide the number of clusters in K-means clustering?

A suitable K needs to be decided by trial and error. This can be decided by plotting SSE for various values of number of clusters. An optimal number of K is where there is an elbow or dip in the graph. Silhouette coefficient (explained in the question below) can also be used instead of SSE to find the optimal number of clusters.

Question 10) What is silhouette coefficient in clustering?

Silhouette coefficient is a cluster evaluation method that combines cohesion and separation measures.

$$\text{Silhouette coefficient} = \frac{b_i - a_i}{\max(a_i, b_i)}$$

a_i = Average distance of data object i from all other objects in its cluster

b_i = Average distance of data object i from other objects that are from a different cluster than i is calculated. b_i is the minimum distance with respect to all the clusters.

The value of the silhouette coefficient will vary from -1 to 1.

A negative silhouette coefficient is undesirable. A value that is close to 1 is desirable, and this is achieved when the value of ai is 0 or close to 0.

The silhouette coefficient of a cluster is obtained by taking the average of all silhouette coefficients of data objects in a cluster. This value is a measure of the goodness of a cluster, which can be used for comparison between clusters.

Decision Trees

Common Interview Questions - Part 1

Question 1) What are the merits and demerits of the decision tree model over the linear regression model?

Merits:

- Decision tree is a non-parametric model (no assumptions about the data), while the regression model is parametric
- The decision tree model can learn non-linear decision boundaries also, which is not possible in regression models such as linear and logistic regressions
- It can be used for both regression and classification problems
- It is easy to visualise (in case of deep trees, visualisation is difficult)

Demerits:

- Getting business insights from the decision tree model is very difficult
- Decision trees are weak classifiers and have to be used with bagging, random forests, and other algorithms to perform efficiently
- Decision trees cannot explain the marginal effects of a variable on the target, i.e. how the target will change if a variable is changed by 1 unit
- Hyperparameter tuning is required in decision trees

Question 2) There are 50 predictors in a dataset. You have built two models on this dataset: bagged decision tree and random forest. Let the number of predictors used on a single split in a bagged decision tree be X and random forest be Y. What can you say about X and Y?

X>=Y. Random forest uses a random subset of the predictors, while the bagged decision trees always use all the predictors on a single split. Also, there may be cases when the random forest uses all the features. In such a case, X and Y will be equal, so overall, X>=Y.

Question 3) Explain the random forests algorithm.

Random forest is an ensemble learning method, which tries to build a multitude of decision tree (CART) models to predict the target variable. It builds many decision trees with random sample and decision variables. If used for regression purposes, it outputs the mean of all the decision trees. In case of a classification problem, the mode of the outputs of the decision trees is the final output.

Note: For details, you can refer to the main content and revise the concepts.

Question 4) Do random forests overfit?

Random forests do not generally overfit. Random forests, being an ensemble learning, try to reduce the variance. But at certain times, due to some useless features present in the data and a large number of trees, it might overfit by a small margin.

Question 5) What is 'random' in the random forest algorithm?

Random forest tries to build many decision trees with different sample and decision variables. The selection of sample and decision variables in a decision tree is made randomly. Hence, it's called random forest.

Question 6) What is meant by bagging?

Bagging refers to Bootstrap Aggregation, which is the technique used to reduce the variance of decision tree models. In the Bagging algorithm, several subsets of the data are randomly chosen with replacement and are fed into separate decision tree models, creating an ensemble of decision tree models. The output is the average of the predictions made by all the decision trees, making it more robust and reducing the variance.

Question 7) What do you mean by boosting?

The idea behind boosting is to create a 'powerful' committee of many weak classifiers. Here, samples are collected from the training data without replacement and are fed into the tree models. The point to note here is that this sampling is not parallel, it is done sequentially; and in each step, more weightage is given to the wrongly classified labels. The final output is the weighted majority vote of all the models trained in this ensemble learning technique. The main aim of the boosting algorithm is to increase the predictive accuracy of the weak classifiers.