Machine Learning Interview Q&A

Q.1 What is curse of dimensionality? How would you handle it?

A.1 The dimensions of a model are decided by the number of features. If a model has large number of features this means it deals with a large dimensional space. A machine learning model can comfortably work with let's say 50-dimensional space, suddenly increasing the dimensions to 500 would be computationally expensive and practically challenging. Learning from 500 different features confuses the model and provides a weird model that is less effective. This problem caused by increased number of features is called "Curse of dimensionality".

Ways to handle "Curse of dimensionality":

There are many ways to reduce the number of features. Simple method to select features which are more important. Feature selection can be done during EDA. Further advanced techniques of feature reduction include PCA, Factor Analysis and Linear Discriminant Analysis.

Q.2 How to find multicollinearity in the dataset?

A.2 Multicollinearity denotes a strong relation between the predictor variable. It can affect the linear regression model and one of the basic assumptions of linear regression is that there should be no multicollinearity in the dataset. There are various ways to detect it:

a. Heatmap: It provides the linear correlation values of all the variables in pictorial presentation.

b. Correlation matrix: It provides the linear correlation values for all variables directly.

c. VIF: Variation inflation factor captures the effect on multicollinearity on the regression coefficients. It is calculated as (1/ 1- R2). A default threshold of 5 is considered. If the VIF value is larger than 5 it indicates high multicollinearity.

Q.3 Explain the different ways to treat multicollinearity?

A.3 Once multicollinearity is depicted for a specific feature in the dataset there are two ways to treat it:

1. Drop the feature with high multicollinearity if it does not have a strong relation with the target variable.

2. In case if that variable is important, it would not be suggested to drop that feature. Feature engineering allows to make use of original features. For example: in housing price project number of rooms and carpet area are high correlated. A new feature: Area per room can be obtained by dividing the two features. It removes multicollinearity from the dataset and also retains the significant information provided by the two variables.

Q.4 Explain Logistic Regression.

A.4 Logistic Regression is a supervised Machine Learning Algorithm used exclusively for classification problems. It works on the concept of linear regression plus sigmoid function so that the numerical output

of regression can be transformed to 0 and 1 with help of sigmoid. Sigmoid transformation allows us to look at the target value as the logarithm of odds ratio, hence it is called “Logistic Regression” or Logs Odds Ratio or Logit Function. Theoretically it is designed to handle binary classification but practically Logistic Regression can work efficiently on multi-class classification problems as well.

Q.5 Why is log loss used in logistic regression?

A.5 Generally the cost function used in linear regression is (1/N)(y – y hat)^2, this is also called Mean Squared Error (MSE). Now for logistic regression since the predicted values are calculated using sigmoid function which is a non-linear function, finding the optimized value of cost function is complicated. Hence to address this issue the cost function used in logistic regression is “Log Loss”. Taking log of corrected probabilities provides a better measure of evaluating the model.

Q.6 What is P value and its significance in Statistical Testing?

A.6 P value is mathematically the probability corresponding to the test statistic. Theoretically it represents the chance of the null hypothesis to be true. Small the P value , lesser is the chance of the null hypothesis and hence higher the chance of “rejecting the null hypothesis”. To decide a threshold level of significance is referred to. If P value is less than level of significance the decision is to reject the null hypothesis. Hence it is a important concept and value for decision making in statistical testing.

Q.7 Why do you use feature selection?

A.7 The data generally provided may contain several features which may or may not be relevant to the problem statement. Unnecessary inclusion of such features may misguide the model and make it more time consuming. Hence it is important to identify which features are important and useful with respect to the given problem statement. This process of selecting the useful features is called feature selection. It helps to produce an effective and efficient model.

Q.8 What are the confidence intervals of the coefficients of linear regression?

A.8 All the coefficients are just estimations based on the sample dataset. Since these are just estimations, how confident one is about these values can be understood by the concept of confidence intervals. For example, if a particular coefficient value is obtained as (B1), a 95% confidence interval would be interpreted as “out of 100 samples 95 will have the true coefficient value as B1”.

Q.9 How do you pick k in k-means?

A.9 k in k-means denotes the number of neighbors to to be referred in order to predict the unknow class. General rules to pick a good value of “k” are as follows:

1. Preferably k must be an odd value to avoid tie
2. if tie still happens increase or decrease the value of k

3. Create an accuracy curve for each k value and look for the segment of the curve where the accuracy value is high and consistent for 2- 3 k values continuously.

Q.10 What are the assumptions of linear regression model?

A.10 The basic assumptions of linear regression model are:

1. The relationship of dependent and independent variables must always be linear
2. No multicollinearity: this implies no two predictor variables must be correlated
3. No autocorrelation: this implies the residuals must not depend on each other.
4. Homoscedasticity: this implies the error term must be same across all predictors.

Q.11 List out data validation techniques.

A.11 Data validation implies the tools and techniques used to ensure to correct and valid data is obtained for model building. A few data validation techniques are: Checking for null values, checking for outliers, visualizing the dataset to trace any unusual values. Other data validation techniques are: data type validation, consistency validation and constraint validation.

Q.12 What is KNN imputation?

A.12 One of the ways of handling null values is KNN imputation. KNN is a ML model known as K nearest neighbors. Accordingly, any unknown value can be best predicted by the observing the values k nearest neighbors. It helps in precise handling of null- values.

Q.13 Why is rotation important in PCA?

A.13 PCA is a machine learning unsupervised model designed to reduce the dimensions of the dataset. It works with the concept that the maximum information is communication by the features with maximum variance. Hence all the datapoints are rotated in the direction of maximum variance. This new direction corresponding the maximum variance is called the first principal component.

Q.14 Differentiate between decision tree and random forest.

A.14 Decision tree and random forest both are supervised machine learning algorithms that can be used for regression and classification both. A Decision tree is a model that is based on splitting the features such that it leads to pure sub nodes and helps in better performance of the model. Selection of best features is done with help of gini or entropy splitting criteria. Random forest on the other hand is a model that works with multiple decision trees being created parallelly. The ensemble technique used is : bootstrapping or bagging using simple random sampling with replacement. It works on the concept that the model's predictive power will increase since it is based on multiple decision tree outputs.

Q.15 How are missing values handled?

A.15 There are several techniques to handle the null values. Filling the null values is called imputation. Depending on the type of feature following are the methods:

Quantitative data: 1. Mean/ median imputation, 2. Random Sample Imputation , 3. End of distribution imputation , 4. Arbitrary value imputation

Categorical data: 1. Mode imputation, 2. Creating a new category to denote null values, 3.KNN imputation 4. Capturing the missing in new feature.

Q.16 How to handle outliers?

Q.16 Outliers are the values with are either too extreme or unusual as compared to the remaining data points. The best ways to identify outliers are as follows:

1. Visualize through boxplots or histograms

2. Z score method: Any z score value less than -3 or greater than +3 is considered as outlier

3. IQR method: In this method upper and lower thresholds are calculated. Any values beyond these thresholds are considered as outliers.

Once the outliers are detected a proper root cause analysis is done to understand if the unusual value is important for model training or is it obtained by any error. Based on that the decision to either keep the unusual value or to remove it from the dataset is taken.

Q.17 Which measure of central tendency is better; mean, median or mode?

A.17 The selection of correct measure of central tendency depends on the type of feature and its distribution. Mean is used for quantitative datasets with normal distributions, Median is used for quantitative datasets with skewed distributions and mode is used for categorical dataset.

Q18 Explain confusion matrix.

A.18 Confusion matrix is one of the most commonly used evaluation metrics for classification models. It comprises of the actual and predicted classes based on which the predictions can be categorized as: True Positive , True Negative, False Positive and False Negative. This helps trace the accuracy rate and the error rate of the model. It also helps in calculating recall and precision of the model.

Q.19 Explain PCA.

A.19 Principal Component Analysis (PCA) is a dimension reduction technique. It works based on the concept that when there are many features, the features with maximum variance provide maximum information about the target variable. Hence the initial feature matrix is first normalized and converted to z scores or a

normalized matrix. Once the normalized matrix is obtained, the corresponding variances and covariances of all pairs of features are calculated. This gives us the variance-covariance matrix. The diagonal elements of this matrix correspond to the variance values. These values are arranged in descending order. The feature with highest variance gives the direction of the first principal component. This direction corresponding to the highest feature is called Eigen Vector and the corresponding variance value is called Eigen Value. The second highest variance corresponds to the second principal component. And like this, various principal components are obtained. All the principal components are orthogonal.

Q.20 Explain K means clustering.

A.20 K means clustering is an unsupervised machine learning algorithm. It works based on the concept of tracing similarities in features and grouping them accordingly. The K here denotes the number of clusters. Clusters is a group of all similar datapoints. A simple way to understand is with the help of an example of a food truck vendor who wants to locate three food trucks in the city. So initial location of the trucks is chosen randomly. The customers go to the truck closest to them. Once the customer segment for each truck is created, the truck observes the center point of that segment and moves in that direction. This results in the change of customers as now few customers might find another truck closer to them. This results in new customer segments. Again the trucks observe the center points of these segments and move accordingly. After few iterations the proper mid points of each cluster are reach and the final customer segments or groups are created. This is exactly how K means algorithm also works.

Q.21 What is K means++?

A.21 K means++ is the updated version of K means. In K means the initially the clusters are created randomly. If these clusters are started from lets say the point of outliers, it may take several iterations to create the final clusters. In order to avoid these scenario undated K means algorithm K means++ is introduced that allows to initiate the clusters more sensibly such that the number of iterations required to create final clusters is much less.

Q.22 How is KNN different from K means?

A.22 KNN is a supervised ML algorithm and K means is an unsupervised algorithm. This implies the dataset provided in KNN is labeled and the aim of this model is to predict the label of the unknown datapoint. In K means algorithm the dataset is not labeled. All the features are plotted on multidimensional space and based on the similarity or proximity of those points clusters are created.

Q.23 If you are asked to predict whether or not a certain company will declare bankruptcy in the next 7 days. {Would you treat this as a classification or a Regression problem}

A.23 The decision of whether to consider it a classification problem or regression problem basically depends on how the data is provide and what is the type o target variable. If the target variable is share price or market value of the company. This is numerical in nature and whenever the target variable is numerical in nature it is a regression problem. Since the original stock market data is provided in this format only ideally it is a regression problem. But suppose if the dataset provided to you is just a few features and the label whether they are bankrupt or not. In that case the target variable is categorical in nature and hence it would become a classification problem.

Q.24 What is Gradient Descent?

A.24 Gradient descent is an optimization technique. It is used in training machine learning and deep learning models. It works on the concept of loss function or cost function. The values of the parameters get initialized randomly, and then with each iteration the model learns and the values of the parameters are altered in the direction that reduces the cost. For example, in linear regression the parameters involved are slope and intercept. These values are initiated randomly and with each step the slope (gradient) is moved in the direction that helps in reducing(descent) the mean squared error value (cost function). The aim is to reach the global minima value of the cost function.

Q.25 What is the learning rate and why do we need to reduce it or increase it?

A.25 Learning rate is known as the step size in the gradient descent algorithm. It is denoted by alpha. It is used in calculated the next slope value as follows:

Slope(next step) = Present slope – (alpha)*(first derivative of slope)

Higher alpha values, result in large step size and does chances of missing the minima increases. Extremely small step size results in several iterations to reach the minima. Hence it is important to have an optimum value of alpha. Hence the value of alpha is increased or decreased as per the behavior of the cost function.

Q.26 What is loss function?

A.26 In simple terms loss function measures the difference between actual and predicted values and guides the model to improve the performance by moving in the direction that leads to reduction in the loss. Suppose actual value is 10 and predicted value is 7, so there is a mismatch of 3 units. This is the error or the loss value of the model. Smaller the loss the better the model. The model continuously iterates the steps to reach the minimum value. Once the minimum error value is reached, the model stops learning as going any further would lead to increase in loss. Loss function to error of one single training , whereas cost function technically refers to average value of the error across all the trainings.

Q.27 What is AUC-ROC?

A.27 AUC- ROC is the Area Under the Curve of Receiver's Operating Characteristics. ROC is a plot between FPR and TPR. If the TPR value is greater than FPR then the model is considered to do a good job. In order to evaluate it mathematically area on the curve of ROC is calculated. Higher the area value better is the model's performance.

Q.28 What is VIF?

A.28 VIF is called Variance Inflation Factor. VIF tells how much the variance in the regression coefficient increases if the predictor variables are related.

It is given by: VIF = 1/(1 – R2).

In a way it depicts the extent of multicollinearity. A value of 5 or above indicates high multicollinearity.

Q.29 What is cross-validation?

A.29 Cross validation is a powerful tool that allows the model enhance its performance during the training stage such that the problem of overfitting can be avoided. It is also called K-fold validation. Here k denotes the number of splits. Ideal value of k lies between 5 to 20. In cross validation the training dataset is split in k (lets say 10 here) equal parts. In each iteration of training one-part acts as test and remaining (k-1) parts act as training dataset. For example, for 10 splits, in first iteration first split acts as test, remaining 9 act as training. In Second iteration second split acts as test and remaining 9 acts as split. And in similar way total 10 iterations are done in which the model gets to be trained and tested and improve from the feedback so obtained. This helps the model to learn from the dataset in a more guided manner and helps in creating a powerful model.

Q.30 What is sigmoid function?

A.30 Sigmoid function is a transformation function that helps to transform the value of a random variable from (– infinity to + infinity) to (0 to 1). It is employed in logistic regression where the predicted variable can take any value on a real number line, but our aim is to convert this variable in 0 and 1, such that it helps classify the two categories. The curve of sigmoid function is S shaped. Formula: (1/ 1 + e^-y).

Q.31 Can sigmoid function be used to handle more than 2 categories?

A.31 Though sigmoid function is designed to handle only two categories at a time, it can practically be used to handle more than 2 categories. For example, in case of two categories there is only one threshold: 0.5. But if there are three categories, it would require two thresholds: 0.33 and 0.67 to classify the three categories. Value with probability less than 0.33 will be considered as class I , probability values between 0.33 to 0.67 will be categorized as class II and the all values greater than 0.67 are categorized as class III. In the similar way sigmoid function can handle multiple class problems easily.

Q.32 Which method is used to split a node in decision tree?

A.32 There are four different methods to decide the splitting feature in a decision tree: Gini , entropy, Chi square score and reduction in variation. Gini denotes the purity of each node. It is calculated based on the sum of square of proportions of each category in each node. Entropy denotes the randomness in each subnode and hence indicates impurity of a node. It is calculated using logarithmic function. Chi square score is a statistical method that makes use of actual counts and expected counts of observations from each class in each sub node. These all methods work fine when the target variable is categorical in nature. The method used when target variable in numerical in “Reduction in Variance”. Accordingly high variance in a subnode denotes high impurity. Hence the split that results in least possible weighted variance is the best split.

Q.33 What is ensembling technique?

A.33 Ensembling is a technique of combining multiple decision trees. Either the decision trees can be created parallelly leading to a bagging ensembling technique, or the decision tress can be created sequentially leading to a boosting ensembling technique. Both techniques are powerful and help in enhancing the models performance to a great extent.

Bagging is the ensembling technique used in Random Forest ML algorithm. Boosting is the ensembling technique used in XGBOOST ML algorithm.

Q.34 What is the difference between linear and logistic regression?

A.34 Linear regression is a model used exclusively for regression problems whereas logistic regression model is exclusively used for classification problem. Linear regression is based on two mathematical approaches: 1) ordinary least square method 2) gradient descent. Logistic regression is based on: gradient descent and sigmoid transformation.

Q.35 How do you decide when to stop splitting the decision tree?

A.35 Decision tree is a greedy algorithm, this implies that it goes on splitting nodes to subnodes till the time it gets the purest end nodes. This results in overfitting hence it is important to apply some constraints to stop the tree from splitting further. There can be many such constraints. Example: maximum number of features to split, maximum depth of decision tree, minimum number of datapoints in decision nodes, minimum number of datapoints in leaf nodes and so on. Any of these constraints can be applied to let the splitting stop and provide an optimum decision tree model.

Q.36 Which clustering technique uses combining of clusters?

A.36 Agglomerative Hierarchical Clustering technique works by combining smaller clusters into larger ones. The nearest data points create small clusters. The nearest small clusters combine to create large clusters and so on the process goes until the desired number of clusters are obtained.

Q.37 Which is the oldest probability distribution?

A.37 Binomial Distribution is one of the oldest distributions. It is a discrete distribution that works on the assumption that the experiment has only two possible outcomes such that each trial is independent of each other with fixed probability of success in each trial and the number of trails is also known.

Q.38 How is random forest different from gradient boosting algorithm given both are tree-based algorithm?

A.38 Random Forest is based on bagging ensemble technique. This implies multiple decision trees are created parallelly. Gradient boosting is based on boosting ensemble technique. This implies multiple decision trees are created sequentially one after the other. The main aim is to pass on the feedback of

previous tree to the next tree so as to learn the error datapoints better and enhance the model performance.

Q.39 How would you build a model to predict credit card fraud?

A.39 Problem statement: to identify if a transaction is fraudulent or not. This implies the target variable here is "categorical" in nature and hence it is a classification problem. Data selection: The key traits that indicate if a transaction is fraudulent or not are: time of transaction/ location of transaction/ system used for transaction/ Number of trials taken to enter correct pin and many more. Once these observations are available a binary classification model can be created using logistic regression or naïve bayes or random forest to train the model and evaluate it. Once the model is created it can evaluation using confusion matrix, recall and precision.

Q.40 How do you derive new features from existing features?

A.40 Deriving appropriate new features depends on the domain and problem statement. Sometimes simple mathematical operations can also help obtain new features. Like mean monthly stock price. Another example is: Directors present age – directors date of joining = directors experience. This gives us a more useful feature in terms of model building. Like this different feature can be created from existing features either by simple transformations or by combining multiple features together.

Q.41 Explain overfitting and what steps you can take to prevent it?

A.41 Overfitting is the scenario of getting low bias and high variance. Low bias implies high accuracy during training stage and low accuracy during testing stage. This happens when the model ends up learning the noise in the training dataset and fails to do correct predictions for the new dataset in form of test dataset. Overfitting can be addressed by applying regularization parameters like lasso and ridge regressions. In this methods a penalty term is added to the cost function and during the optimization process the aim is to minimize the penalty so that there is no overfitting.

Q.42 Why does SVM need to maximize the margin between support vectors?

A.42 SVM works with the concept of segregating the categories in higher dimension space. The wider the gap between two categories the better is the separation and lesser are the chances of misclassification. Hecne the model aims to maximize the margin between the support vectors. To indicate this gamma is used with a low value in the model. Gamma is a hyperparameter used in SVM that lets the model decide whether to go with a large margin (far support vector) or small margin (near support vector). A large gamma value indicates small margin and vice versa.

Q.43 A large margin may lead to misclassification in SVM, how can it be addressed?

A.43 A low gamma value ensures a large margin between support vectors but at the same time it lets few datapoints to be misclassified. If more values are misclassified this will result in lower accuracy of the

model. Hence this problem must be handled by introducing a penalty term call regularization parameter C. This penalty tells the model that misclassifications cannot be tolerated. A low value of C indicates liberal attitude towards misclassifications and a high value of C indicates low tolerance for misclassification.

Q.44 Explain about machine learning?

A.44 Machine Learning is a part of artificial intelligence and computer science. The aim is helping the machine learn from the underlying trends in the dataset and help answer various problem statements like predicting sales, recommending right content to viewer, predicting recovery path of patients, responding right information by chatbots and so on. Various ML models are designed based on different mathematical and statistical concepts that help solve various problem statements. These models can be divided as supervised, unsupervised and reinforcement models.

Q.45 What is the difference between standard scaler and normal scaler?

Q.45 Standard scaler is a scaling technique that converts the raw scores into z scores ranging between -3 to +3 . This method of scaling is best suited when the features follow normal distribution. Normal scaler is a scaling technique that tends to change the shape or distribution of the data. It converts the raw datapoints to a scale of 0 and 1.

Formula of standard scaler = (X – Xmean) / Xstd

Formula of normal scaler = (X – Xmean)/(Xmax – Xmin)

Q.46 What is multiple linear regression?

A.46 Multiple linear regression is a statistical technique that helps develop a model between target variable and multiple predictor variables such that the model follows linear relationship. The aim is to find the best fitted line for all the data points such that the deviation of all actual points from the predicted points on the line are least. For this it calculates the residual of actual minus the predicted values. The sum of squares of residuals for all possible lines is compared and the line with the least possible sum of squared residuals (SSR) is selected as the best fit line of regression. This method followed by multiple linear regression is called ordinary least square (OLS) method.

Q.47 R2 vs Adj R2, which one to prefer?

A.47 R square talks about the percentage of variation in Y that is explained by the model. If more percentage of changes in Y can be explained by the model, that the model is considered to be a good fit.

In a model, when the predictor variables are increased, this causes the R2 value to increase implying that increasing the count of variables leads to a good model, which is not true. Hence when the number of variables is increased, a better measure to evaluate the model's performance is Adjusted R square. In adjusted R square, the value of R square is adjusted for the number of predictor variables. Now as the variable count increases, the value of adjusted R square will increase only when the added variable really adds to the model's explanation power.

Q.48 What is bias and variance and how the trade-off is achieved?

A.48 Bias is the error during training stage and variance is the error during testing stage. Bias happens when the model is too simple and variance happens when the model is too complex. The aim is to ensure that the model complexity is such that both bias and variance are low. In order to achieve this regularization techniques like ridge, lasso and elasticnet are used.

Q.49 How much time does SVM takes to complete if 1 iteration takes 10 second for 1st class and there are 4 classes? What is Kernel in SVM?

A.49 SVM has a time complexity of O (N square). This implies if there are 4 classes the time taken will be 4^2 = 16 times of 10 = 160 seconds.

Kernel is SVM are the transformation functions that transform the dataset into higher dimension space such that the categories are separable. There are different types of kernels like radial basis function (rbf), linear function, polynomial function and so on. These kernels define the transformational relation. By default, the function used is rbf.

Q.50 what is data leakage? How can data leakage be avoided?

A.50 Machine Learning models work on the concept of splitting the data. The first split called the training dataset is used to train the model. This is the dataset that the model is exposed to. The second split is the called the testing dataset. This split plays a pivotal role in machine learning as this dataset is not seen by the model and it helps us to evaluate the model's true performance on unseen data. If by any chance the model gets to see the testing dataset during the training stage this will leak the unseen data to the model. This is called "data leakage". Now when the model used to do predictions for testing dataset it is going to easily predict it, showing a very optimistic and biased performance.

Data leakage can be solved by ensuring that by no means the model gets to see the test dataset. While preprocessing the features it must be ensured that all the calculations are done based on training dataset and the same transformation is applied on training and testing dataset.

Best Wishes

Amrita Panjwani