Question. You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)

Processing high-dimensional data on a limited memory machine is a strenuous task,

1. Since we have lower RAM, we should close all other applications on our machine, including the web browser, so that most of the memory can be used.

2. We can randomly sample the data set. This means we can create a smaller data set, let’s say, having 1000 variables and 300000 rows, and do the computations.

3. To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we’ll use correlation. For categorical variables, we’ll use the chi-square test.

4. Also, we can use PCA and pick the components that can explain the maximum variance in the data set.

5. Using online learning algorithms like Vowpal Wabbit is a possible option.

6. Building a linear model using Stochastic Gradient Descent is also helpful.

7. We can also apply our business understanding to estimate which predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in a significant loss of information.

Question What Are Some Methods of Reducing Dimensionality?

We can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

Question What is Principal Component Analysis?

Principal Component Analysis or PCA is a multivariate statistical technique that is used for analyzing quantitative data. The objective of PCA is to reduce higher dimensional data to lower dimensions, remove noise, and extract crucial information such as features and attributes from large amounts of data.

Question Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?

Yes, Rotation helps simplify the interpretation of the components by aligning them with the axes that maximize the variance of the data. This makes the components easier to interpret. that’s the motive of doing PCA where we aim to select fewer components (than features) that can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn’t change, it only changes the actual coordinates of the points.

Without rotation, the components may be correlated with each other and we’ll have to select more components, leading to redundancy and difficulty in interpretation.

Question. You are given a data set on cancer detection. You’ve built a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?

As this is hospital data, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting the majority class correctly, but our class of interest is the minority class (4%) which is the people who got diagnosed with cancer. Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), and F measure to determine the class-wise performance of the classifier. If the minority class performance is found to be poor, we can undertake the following steps:

1. We can use undersampling and oversampling to make the data balanced. 2. We can alter the prediction threshold value by doing probability calibration and finding an optimal threshold using the AUC-ROC curve.

3. We can assign weight to classes such that the minority classes get larger weight. 4. We can also use anomaly detection.

Question. Why is naive Bayes so ‘naive’?

Naive Bayes is so ‘naive’ because it assumes that all of the features in a data set are equally important and independent. As we know, these assumptions are rarely true in real-world scenarios.

For example, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

Question. Explain prior probability, likelihood, and marginal likelihood in the context of the naive Bayes algorithm.

Prior probability is nothing but, the proportion of dependent (binary) variables in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chance that any new email would be classified as spam.

Likelihood is the probability of classifying a given observation as 1 in the presence of some other variable. For example: The probability that the word ‘FREE’ is used in a previous spam message is likelihood. Marginal likelihood is, the probability that the word ‘FREE’ is used in any message.

Question . You are working on a time series data set. Your manager has asked you to build a high-accuracy model. You start with the decision tree algorithm since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than the decision tree model. Can this happen? Why?

Time series data is known to possess linearity. On the other hand, a decision tree algorithm is known to work best to detect non–linear interactions. The reason why the decision tree failed to provide robust predictions is because it couldn’t map the linear relationship as well as a regression model did. The linear regression model can provide robust prediction given the data set satisfies its linearity assumptions.

Question . You are assigned a new project which involves helping a food delivery company save more money. The problem is, the company’s delivery team aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?

You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine-learning fundamentals.

This is not a machine-learning problem. This is a route optimization problem. A machine learning problem consists of three things:

1. There exists a pattern.

2. You cannot solve it mathematically (even by writing exponential equations). 3. You have data on it.

Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

Question. You learned that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

Low bias occurs when the model’s predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like a great achievement, do not forget, that a flexible model has no generalization capabilities. It means, that when the model is tested on unseen data, it gives disappointing results.

In such situations, we can use a bagging algorithm (like a random forest) to tackle high-variance problems. Bagging algorithms divide a data set into subsets made with repeated randomized sampling. Then, these samples are used to generate a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).

Also, to combat high variance, we can:

1. Use the regularization technique, where higher model coefficients get penalized, hence lowering model complexity, feature selection, and ensemble learning.

2. Use top n features from the variable importance chart. Maybe, with all the variables in the data set, the algorithm is having difficulty in finding the meaningful signal.

Question. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?

PCA assumes that the variables are uncorrelated. Correlated variables can introduce multicollinearity, which can affect the accuracy and interpretability of the PCA results. Removing correlated variables helps reduce redundancy and ensures that the principal components capture the maximum amount of variance in the data. By removing correlated variables, we can avoid the issue of one variable dominating the principal components and potentially distorting the results.

Question . After spending several hours, you are now anxious to build a high-accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of the models could perform better than the benchmark score. Finally, you decided to combine those models. Though ensembled models are known to return high accuracy, you are unfortunate. Where did you miss?

As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. However, these learners provide superior results when the combined models are uncorrelated. Since we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, that All the models provide the same information. GBM models may have high variance, leading to overfitting and poor generalization performance

For example: If model 1 has classified User1122 as 1, there are high chancethat s that model 2 and model 3 would have done the same, even if their actual value is 0. Therefore, ensemble learners are built on the premise of combining weak uncorrelated models to obtain better predictions.

Question. What are some real-life applications of clustering algorithms?

The [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) technique can be used in multiple domains of data science like image classification, customer segmentation, and recommendation engine. One of the most common use is in market research and customer segmentation which is then utilized to target a particular market group to expand the businesses and profitable outcomes.

Question. How is kNN different from K-means clustering?

The fundamental difference between these algorithms is, that k-means is unsupervised and kNN is supervised. K-means is a clustering algorithm. kNN is a classification (or regression) algorithm.

K-means algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to the unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number ) surrounding neighbors. It is also known as lazy learning because it involves minimal training of the model. Hence, it doesn’t use training data to generalize on unseen data sets.

Question . How is True Positive Rate and Recall related? Write the equation.

TPR is defined as the ratio of true positive predictions to the total number of actual positive instances in the dataset. Recall, on the other hand, is the proportion of true positive predictions out of all the positive instances. Mathematically, TPR and Recall are the same and can be calculated using the following equation: Recall = TPR = TP / (TP + FN).

Question. You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, you remove the intercept term, and your model R² becomes 0.8 from 0.3. Is it possible? How?

Yes, it is possible. We need to understand the significance of intercept in a regression model. The intercept term in a regression model represents the average value of the dependent variable when all independent variables are zero By removing the intercept term, the model is forced to fit the data through the origin, which can lead to a better fit if the data has a natural origin.

Question. After analyzing the model, your manager has informed us that your regression model is suffering from multicollinearity. How would you check if he’s true? Without losing any information, can you still build a better model?

To check multicollinearity, we can create a correlation matrix to identify & remove variables correlating 75% (deciding a threshold is subjective). In addition, we can calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.

However, removing correlated variables might lead to a loss of information. To retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variables so that the variables become different from each other. However, adding noise might affect the prediction accuracy, hence this approach should be carefully used.

Question. When is Ridge regression favorable over Lasso regression?

Ridge regression is favorable over Lasso regression when dealing with multicollinearity in the data. Ridge regression is more suitable when the sample size is larger than the number of variables (p < n). As it does not shrink coefficients to precisely zero, allowing for a more stable and less biased estimation. Regression is less sensitive to changes in the ridge parameter, making it more robust in sub-optimal parameter selection. Ridge regression is preferred when the goal is prediction accuracy rather than model interpretability.

Question. The rise in global average temperature led to a decrease in number of pirates around the world. Does that mean that a decrease in a number of pirates caused climate change?

After reading this question, you should have understood that this is a classic case of “causation and correlation”. No, we can’t conclude that a decrease in number of pirates caused climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.

Therefore, there might be a correlation between the global average temperature and several pirates, but based on this information we can’t say that pirated died because of the rise in global average temperature. it's not a meaningful explanation for the causes of climate change

Causation: Causation refers to a cause-and-effect relationship between two variables, where one variable directly influences or causes a change in the other variable.

Question. While working on a data set, how do you select important variables? Explain your methods.

One method for selecting important variables is through the use of statistical techniques such as regression analysis. Regression analysis can help identify variables that have a significant impact on the outcome variable. Another approach is to use feature selection algorithms, such as Recursive Feature Elimination (RFE) or Lasso regression, which can automatically select the most relevant variables based on their predictive power. Correlation analysis and data visualization can help identify variables that are highly correlated with the outcome.

Question. What is the difference between covariance and correlation?

Covariance measures the direction and strength of the linear relationship between two variables. It indicates how changes in one variable are associated with changes in another variable. Correlation, on the other hand, is a standardized measure of the linear relationship between two variables. It ranges from -1 to 1 and provides information about both the direction and strength of the relationship. Covariance is affected by the scale of the variables, whereas correlation is not. Correlation is a more commonly used measure because it provides a standardized way to compare the strength and direction of relationships between variables, regardless of their units of measurement.

Question . Is it possible to capture the correlation between continuous and categorical variables? If yes, how?

Yes, we can use the ANOVA (analysis of covariance) technique to capture the association between continuous and categorical variables.

Question. Both being tree-based algorithms, how is random forest different from Gradient boosting algorithm (GBM)?

The fundamental difference is, that random forest uses bagging techniques to make predictions. GBM uses boosting techniques to make predictions.

In the bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done in parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding round. This sequential process of giving higher weights to misclassified predictions continues until a stopping criterion is reached.

Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy by reducing both bias and variance in a model.

Question. Running a binary classification tree algorithm is the easy part. Do you know how a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?

A classification tree makes decisions based on the Gini Index and Node Entropy. In simple words, the tree algorithm finds the best possible feature that can divide the data set into the purest possible children nodes.

The Gini index says, that if we select two items from a population at random then they must be of the same class, and the probability for this is 1 if population is pure. We can calculate Gini as follows:

1. Calculate Gini for sub-nodes, using the the formula sum of the square of probability for success and failure (p^2+q^2).

2. Calculate Gini for the split using the weighted Gini score of each node of that split Entropy is the measure of impurity as given by (for binary class):



Here p and q are the probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when both the classes are present in a node at 50% – 50%. Lower entropy is desirable.

Question . You’ve built a random forest model with 10000 trees. You got delighted after getting a training error of 0.00. But, the validation error is 34.23. What is going on? Haven’t you trained your model perfectly?

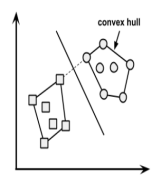
The model has overfitted. Training error 0.00 means the model has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this model was run on the unseen sample, it couldn’t find those patterns and returned a prediction with a higher error. In a random forest, it happens when we use a larger number of trees than necessary. Hence, we should tune the number of trees using cross-validation to avoid these situations.

Question . You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS a bad option to work with? Which techniques would be best to use? Why?

In such high dimensional data sets, we can’t use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

To combat this situation, we can use penalized regression methods like lasso, LARS, and ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

Other methods include subset regression and forward stepwise regression.



Question . What is a convex hull? (Hint: Think SVM)

In the case of linearly separable data, the convex hull represents the outer boundaries of the two groups of data points. Once a convex hull is created, we get a maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line that attempts to create the greatest separation between two groups.

Question . We know that one hot encoding increases the dimensionality of a data set. But, label encoding doesn’t. How?

Using one hot encoding, the dimensionality/features in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let’s say we have a variable ‘color’. The variable has 3 levels namely Red, Blue, and Green. One hot encoding ‘color’ variable will generate three new variables as Color. Red, Color.Blue and Color.Green containing 0 and 1 value.

In label encoding, the levels of categorical variables get encoded as 0 and 1, so no new variable is created. Label encoding is mainly used for binary variables.

Question. What cross-validation technique would you use on a time series data set? Is it K-fold or LOOCV?

Neither. In the time series problem, k fold can be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validating past years, which is incorrect. Instead, we can use a forward chaining strategy with 5 fold as shown below:

● fold 1: training [1], test [2]

● fold 2: training [1 2], test [3]

● fold 3: training [1 2 3], test [4]

● fold 4: training [1 2 3 4], test [5]

● fold 5: training [1 2 3 4 5], test [6]

Question What is Cross-Validation?

Cross-validation in Machine Learning is a statistical resampling technique that uses different parts of the dataset to train and test a machine learning algorithm on different iterations. Cross-validation aims to test the model’s ability to predict a new set of data that was not used to train the model. Cross-validation avoids the overfitting of data. K-fold cross-validation is the most popular resampling technique that divides the whole dataset into K sets of equal sizes. Mainly used in backgrounds where the objective is forecast, and one wants to estimate how accurately a model will accomplish in practice.M

Question. You are given a data set consisting of variables having more than 30% missing values. Let’s say, that out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?

We can deal with them in the following ways:

1. Assign a unique category to missing values, who knows the missing values might decipher some trend

2. We can remove them blatantly.

3. Or, we can sensibly check their distribution with the target variable, and if found any pattern we’ll keep those missing values and assign them a new category while removing others.

Question. ‘People who bought this, also bought…’ recommendations seen on amazon is a result of which algorithm?

The basic idea for this kind of recommendation engine comes from collaborative filtering.

Collaborative Filtering algorithm considers “User Behavior” for recommending items. They exploit the behavior of other users and items in terms of transaction history, ratings, selection, and purchase information. Other user's behavior and preferences over the items are used to recommend items to new users. In this case, the features of the items are not known.

Question. What is data leakage and how can we identify it?

If there is a high correlation between the target variable and the input features then this situation is called data leakage. This is because when we train our model with that highly correlated feature then the model gets most of the target variable’s information in the training process only and it has to do very little to achieve high accuracy. In this situation, the model gives pretty decent performance both on the training as well as the validation data but as we use that model to make actual predictions then the model’s performance is not up to the mark. This is how we can identify data leakage.

Question. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

Since the data is spread across the median, let’s assume it’s a normal distribution. We know, that in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Question. What is the Central Limit Theorem and why is it important?

The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. In simpler terms, if you take repeated random samples from any population, the distribution of the sample means will be approximately normal, even if the original population distribution is not normal. It is important because it allows us to make inferences about a population based on a sample, even when the population distribution is unknown or non-normal.

Question. What is a statistical interaction?

It refers to a situation where the effect of one variable on a dependent variable differs depending on the level of another variable. In other words, the relationship between the independent variable and the dependent variable changes depending on the value of another variable. For example, imagine studying the effects of both exercise and diet on weight loss. If there's a statistical interaction between exercise and diet, it means that the impact of exercise on weight loss depends on the individual's diet, and vice versa.

Question. What is selection bias?

Question. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls? ,

P(Having two girl given one is a girl) 1\2

Selection bias occurs when the sample used in a study does not represent the population it is intended to represent, leading to systematic differences between the sample and the population. This bias can arise when certain individuals or groups are more likely to be included or excluded from the sample, leading to results that may not accurately reflect the true characteristics or relationships within the population.

Question. What do you understand by the term Normal Distribution?

The mean, median, and mode are all equal and located at the center of the distribution. The distribution is symmetric around the mean, with the same proportion of data falling on either side of the mean. The standard deviation determines the spread or width of the distribution. Approximately 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations. The tails of the distribution extend infinitely in both directions but asymptotically approach the x-axis without ever touching it.

Question. What is the difference between Point Estimates and Confidence Interval?

A point estimate provides a single value as an estimate of the population parameter and does not convey the uncertainty associated with the estimate, a confidence interval provides a range of values within which the true parameter value is likely to lie, along with a measure of the level of confidence associated with that range. Confidence intervals are often preferred over point estimates alone because they convey both the estimate of the parameter and the uncertainty associated with it.

Question. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. A cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements. For example, a researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Question. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is the equal probability method.

Question. What do you understand by Type I vs Type II error?

Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we accept it, also known as ‘False Negative’.

In the context of the confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

Question. You are working on a classification problem. For validation purposes, you’ve randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?

In the case of a classification problem, we should always use stratified sampling instead of random sampling. Random sampling doesn’t take into consideration the proportion of target classes. On the other side, stratified sampling helps to maintain the distribution of target variables in the resultant distributed samples.

Question. You have been asked to evaluate a regression model based on R², adjusted R², and tolerance. What will be your criteria?

Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of the percent of variance in a predictor which cannot be accounted for by other predictors. Large values of tolerance are desirable.

We will consider adjusted R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. However, adjusted R² would only increase if an additional variable improves the accuracy of the model, otherwise stays the same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to stock market data where lower adjusted R² implies that the model is not good.

Question. In k-means or kNN, we use Euclidean distance to calculate the distance between nearest neighbors. Why not Manhattan distance?

We don’t use Manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since the data points can be present in any dimension, euclidean distance is a more viable option.

Example: In a chess board, the movement made by a bishop or a rook is calculated by Manhattan distance because of their respective vertical & horizontal movements.

Question. Explain machine learning to me like a 5-year-old.

It’s simple. It’s just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bent position. The next time they fall down, they feel pain. They cry. But, they learn ‘not to stand like that again’. To avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

Q35. I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?

1. Since logistic regression is used to predict probabilities, we can use the AUC-ROC curve along with the confusion matrix to determine its performance.

2. Also, the analogous metric of adjusted R² in logistic regression is AIC. AIC is the measure of fit that penalizes the model for the number of model coefficients. Therefore, we always prefer the model with minimum AIC value.

3. Null Deviance indicates the response predicted by a model with nothing but an intercept. The lower the value, the better the model. Residual deviance indicates the response predicted by a model on adding independent variables. The lower the value, the better the model.

Question. Considering the long list of machine learning algorithms, given a data set, how do you decide which one to use?

The choice of algorithm solely depends on the type of data. If you are given a data set exhibiting linearity, then linear regression would be the best algorithm. If you are given to work on images and audio, then a neural network would help you to build a robust model.

If the data comprises nonlinear interactions, then a boosting or bagging algorithm should be the choice. If the business requirement is to build a model that can be deployed, then we’ll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM, etc.

In short, there is no one master algorithm for all situations.

Question. Do you suggest that treating a categorical variable as a continuous variable would result in a better predictive model?

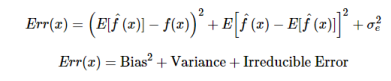
For better predictions, the categorical variable can be considered as a continuous variable only when the variable is ordinal.

Question. When does regularization become necessary in Machine Learning?

Regularization becomes necessary when the model begins to overfit/underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce the cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

Question. What do you understand by the Bias-variance trade-off?

Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand. Variance: Variance is the error introduced in your model due to a complex machine learning algorithm, your model learns noise also from the training data set and performs badly on the test data set. It can lead to high sensitivity and overfitting. Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

Question OLS is to linear regression. The maximum likelihood is logistic regression. Explain the statement.

OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words, Ordinary least squares (OLS) is a method used in linear regression that approximates the parameters resulting in a minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.

Question What is the AUC-ROC curve?

It is an important model metric for evaluating binary classifiers and can be used to compare the performance of different classifiers.

Question How is the ROC curve constructed? What are the axes?

The Receiver Operating Characteristic (ROC) curve is constructed by plotting the True Positive Rate Sensitivity(TP/TP\*FN) against the False Positive Rate Specificity(FP/FP\*TN) at various classification thresholds. The x-axis of the ROC curve represents the False Positive Rate (FPR), which is calculated as the ratio of false positive predictions to the total number of actual negative instances. The y-axis of the ROC curve represents the True Positive Rate (TPR), which is calculated as the ratio of true positive predictions to the total number of actual positive instances.

AUC (Area Under the Curve) in the context of the ROC curve represents the measure of the overall performance or accuracy of a binary classifier.

Question. Can the AUC-ROC curve be used to compare models with different thresholds? If yes, how?

Yes, the AUC-ROC curve can be used to compare models with different thresholds. A higher AUC value indicates a better-performing model, regardless of the specific threshold chosen. Therefore, even if two models have different thresholds, their AUC values can still be compared to determine which model is more effective in classification.

Question. What do you understand by the F1 score?

The F1 score is a metric that combines both Precision and Recall. It is also the weighted average of precision and recall.

The F1 score can be calculated using the below formula:

F1 = 2 \* (P \* R) / (P + R)

The F1 score is one when both Precision and Recall scores are one.

Question. What is a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the binary classifier. Various measures, such as error rate, accuracy, specificity, sensitivity, precision, and recall are derived from it. Confusion Matrix A data set used for performance evaluation is called a test data set. It should contain the correct labels and predicted labels. A binary classifier predicts all data instances of a test data set as positive or negative. This produces four outcomes. True-positive(TP) — Correct positive prediction 2. False-positive(FP) — Incorrect positive prediction 3. True-negative(TN) — Correct negative prediction 4. False-negative(FN) — Incorrect negative prediction

Basic measures derived from the confusion matrix 1. Error Rate = (FP+FN)/(P+N) 2. Accuracy = (TP+TN)/(P+N) 3. Sensitivity(Recall or True positive rate) = TP/P 4. Specificity(True negative rate) = TN/N 5. Precision(Positive predicted value) = TP/(TP+FP) 6. F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b²PREC+REC) where b is commonly 0.5, 1, 2.

Question. How can you conclude about the model’s performance using the confusion matrix?

Confusion matrix summarizes the performance of a classification model. In a confusion matrix, we get four types of output (in case of a binary classification problem) which are TP, TN, FP, and FN. As we know that there are two diagonals possible in a square, and one of these two diagonals represents the numbers for which our model’s prediction and the true labels are the same. Our target is also to maximize the values along these diagonals. From the confusion matrix, we can calculate various evaluation metrics like accuracy, precision, recall, F1 score, etc.

Question. What is TF/IDF vectorization?

TF–IDF is short for term frequency-inverse document frequency, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Question. Explain SVM.

Question. What are the support vectors in SVM?

Support vectors are the data points that lie closest to the decision boundary or hyperplane. These support vectors play a crucial role in defining the decision boundary and determining the classification of new data points. In SVM, the goal is to find the hyperplane that maximally separates the different classes while minimizing the margin or distance to the nearest data points. During the training process of SVM, only the support vectors are used to construct the decision boundary, making SVM memory-efficient and computationally efficient.

Question. What are Entropy and Information gain in the Decision tree algorithm?

Entropy is used to determine the best attribute to split the data on. The goal is to minimize entropy and create homogeneous subsets of data. Where information gain is used to evaluate the effectiveness of a particular attribute in splitting the data. It measures the reduction in entropy achieved by splitting the data based on that attribute. The attribute with the highest information gain is chosen as the splitting criterion. Entropy and information gain are closely related. Information gain is calculated by subtracting the weighted average of the entropies of the child nodes from the entropy of the parent node. It represents the amount of information gained by splitting the data on a particular attribute.

Question. What is pruning in Decision Tree?

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. So, when we remove subnodes of a decision node, this process is called pruning or opposite process of splitting.

Question. What is Selection Bias?

Selection Bias is a statistical error that brings about a bias in the sampling portion of the experiment. This, in turn, causes more selection of the sampling portion than other groups, which brings about an inaccurate conclusion.

Question. What is stemming?

Stemming is a normalization technique that removes any affix joined to a word, leaving it in its base state. It makes text easier to process. It is commonly used in information retrieval, an important step in text pre-processing and text mining applications. Stemming can be used in various NLP tasks such as text classification, information retrieval, and text summarization.

Question. What is Lemmatization?

Lemmatization is a normalization technique that converts a word into a lemma form, or the root word, which is not the stem word. It is a process in which a word is reduced to its base form, but not similar to stemming; it considers the context of the word and produces a valid word. Lemmatization is quite difficult compared to stemming because it requires a lot more knowledge about the structure of a language; it's a much more intensive process than just trying to set up a heuristic stemming algorithm. Lemmatization is often used in natural language processing (NLP) applications to improve text analysis and feature extraction.

Question. What are the differences between over-fitting and under-fitting?

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data. Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Question. How to combat Overfitting and Underfitting?

To combat overfitting: 1. Add noise 2. Feature selection 3. Increase training set 4. L2 (ridge) or L1 (lasso) regularization; L1 drops weights, L2 no 5. Use cross-validation techniques, such as k-fold cross-validation 6. Boosting and bagging 7. Dropout technique 8. Perform early stopping 9. Remove inner layers To combat underfitting: 1. Add features 2. Increase the time of training

Question. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing, or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of the eigenvector or the factor by which the compression occurs.

Question. How Can You Choose a Classifier Based on a Training Set Data Size?

* Simple algorithms like Naive Bayes, Logistic Regression, or Decision Trees can work well with small datasets. These algorithms are less prone to overfitting on small datasets and provide good performance. Algorithms like Support Vector Machines (SVM), Random Forests, or Gradient Boosting Machines (GBM) can handle medium-sized datasets effectively. These algorithms are more robust and can capture complex patterns in the data, but they may require more computational resources and tuning. Deep learning algorithms, such as Convolutional Neural Networks (CNNs) for image data or Recurrent Neural Networks (RNNs) for sequential data, can handle large datasets effectively. Deep learning algorithms can automatically learn hierarchical representations of the data, but they require substantial computational resources, especially for training on large datasets.

Question. How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choosing an algorithm for a classification problem, you can follow these guidelines. If accuracy is a concern, test different algorithms and cross-validate them. If the training dataset is small, use models that have low variance and high bias. If the training dataset is large, use models that have high variance and little bias

Question. What are the assumptions you need to take before starting with linear regression?

There are primarily 5 assumptions for a Linear Regression model:

* Linear relationship: if x goes up y should go up. /down/
* No or little multicollinearity: means features should not be correlated with each other.
  + How to check multicollinearity using heatmap.
* Residual/Error normality: our residual plot should be normal
* No auto-correlation: This means there should be no relationship between all errors/residuals
* Homoscedasticity: Having the same scatter, meaning when we plot the error rate.

Question. Explain linear regression.

Question. What are the different methods to split a tree in a decision tree algorithm?

Gini Index: The Gini index measures the impurity of a node by calculating the probability of misclassifying a randomly chosen element. The split with the lowest Gini index is chosen as the best split.

Information Gain: Information gain measures the reduction in entropy or uncertainty after a split. It calculates the difference between the entropy of the parent node and the weighted average of the entropies of the child nodes. The split with the highest information gain is selected.

Gain Ratio: The gain ratio is an extension of information gain that takes into account the number of branches a split generates. It penalizes splits that create a large number of branches. The split with the highest gain ratio is chosen.

Chi-Square: The chi-square test is used to determine if there is a significant difference between the expected and observed frequencies of a categorical variable. It measures the independence between the target variable and the feature being split. The split with the highest chi-square value is selected.

Reduction in Variance: This method is used for regression problems. It calculates the reduction in variance after a split and selects the split that maximizes the reduction.

Question: What is Ensemble learning?

Ensemble learning is a combination of the results obtained from multiple machine learning models to increase the accuracy for improved decision-making.

Example: A Random Forest with 100 trees can provide much better results than using just one decision tree.

Question. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

In the gradient descent algorithm train our model on the whole dataset at once. But in Stochastic Gradient Descent, the model is trained by using a mini-batch of training data at once. If we are using SGD then one cannot expect the training error to go down smoothly. The training error oscillates but after some training steps, we can say that the training error has gone down. Also, the minima achieved by using [GD](https://www.geeksforgeeks.org/gradient-descent-algorithm-and-its-variants/) may vary from that achieved using the SGD. It is observed that the minima achieved by using SGD are close to GD but not the same.

Question. Why removing highly correlated features are considered a good practice?

When two features are highly correlated, they may provide similar information to the model, which may cause overfitting. If there are highly correlated features in the dataset then they unnecessarily increase the dimensionality of the feature space and sometimes create the problem of the curse of dimensionality. If the dimensionality of the feature space is high then the model training may take more time than expected, it will increase the complexity of the model and chances of error. This somehow also helps us to achieve data compression as the features have been removed without much loss of data.

Supervised Learning Algorithms

Naive Bayes: Naive Bies is a probabilistic machine learning algorithm based on Bies' theorem, with a "naive" assumption of independence between features. despite its simplicity and the naive assumption, Naive Baes is often surprisingly effective, especially in text classification tasks. It's widely used for spam filtering, sentiment analysis, and document categorization. Here are the key components and steps involved in Naive Baes.

Types of Naive Bayes Classifiers

Gaussian Naive Bayes: Assumes that continuous features follow a Gaussian distribution.

Multinational Naive Bayes: Suitable for discrete data, often used in text classification where features represent word counts or term frequencies.

Bernoulli Naive Bayes: Appropriate for binary features, typically used in text classification with binary feature vectors.