Machine Learning

Q1 Define Bias and Variance

Bias

* It is simply defined as inability of model because of that there is some difference or error occuring b/w the model's predicted value and actual value.
* These difference b/w actual or expected values and the predicted values are known as error or bias error.
* $\text{Bias}(\hat{Y}) = E(\hat{Y}) - Y$

Variance

* Variance is the measure of spread in data from its mean position.
* It is amount by which the performance of a predictive model changes when it is trained on different subsets of training data.
* $\text{Variance} = E\left[(\hat{Y} - E(\hat{Y})^2\right]$

Q2 What is structure of learning

Q3 How to make predictions in linear regression

* Making prediction is as simple as solving the equation for a specific inputs.
* Let's make this concrete with an example. Imagine we are predicting weight(Y) from height(X).
* Our linear regression model representation for this problem would be

$$Y = B0 + B1(X)$$

$$Weight = B0 + B1(Height)$$

Q4 What is No-Free lunch theorm?

→ This theorm states that there is no one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem, so it is common in ML to try multiple models and find one that works best for a particular problem.

Q5 What do you mean by Stopping criterion in CART?

→ The most frequent ~~stop~~ halting method is to utilize a min amount of training data allocated to every leaf node. If the count is smaller than the specified threshold, the split is rejected and also the node is considered the last leaf node.

Q6 What are the libraries used to do visualization

* Matplotlib
* plotly
* Altair
* Seaborn
* GGplot
* Geoplotlib

Q7 Explain framework of inductive learning.

* Training example
→ A sample from X, including its o/p from target function.

* Target function
→ A mapping function f from X to f(X)

* Hypothesis
→ Approximation of f, candidate function.

* Classifier
→ learning programs outputs a classifier that is used to classify

* learner
→ process that creates the classifier.

Q8 Define Training set & Test set

Training Set

* It is used to fit the model to learn the parameters of the model
* larger in size as compared to test sets
* slower on larger datasets but job can be run in parallel using multiprocessing

Test Set

* It is used to test whether model can generalize well on unseen data
* Smaller in size as compared to train set
* Faster than train sets

Q9 Define overfiting & underfitting

Overfitting

-> It occurs when our ML model tries to cover all the data points or more than the required data points present in given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.

-> The overfitted model has high variance & low bias.

Underfitting

-> It occurs when our ML model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an eagerly stage, due to which the model may not learn enough from the training data. As a result it may fail to find the best fit of the dominant trend in the data.

-> The underfitted model has low variance and high bias.

-> It can neither model the training data nor generalize to new data.

Q10 What is aggregation?

* Aggregation is the process to reduce the many measurements into a few values/statistics
* You can do this aggregation in many different ways, the mean & 95% confidence interval is just one of many options to summarize the several measurements.

Q11 Give any 3 applications of ML?

* Fraud detection
* Speech recognition
* Stock market analysis & forecasting

Q12 Define Gini ~~cost~~ cost.

* It is a proportion of impurity or inequality in statistical and monetary settings.
* In ML, Gini cost is utilized as an impurity measure in decision tree algorithms for classification tasks.
* It is calculated by subtracting the sum of the squared probabilities of each class from one.

Q13 What do you mean by Gradient Descent?

* When there are one or more inputs you can use a process of optimizing the values of coefficients by iteratively minimizing the error of the model on your training data. This operation is called Gradient Descent and works by starting with zero values for each coefficient.
* When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.
* Gradient descent is often taught using a linear regression model because it is relatively straightforward to understand. In practise, it is useful when you have a very large dataset either in the no. of rows or the no. of columns that may not fit into memory.

Q14 Explain limit overfitting. How to handle the sequence?

* Both overfitting and underfitting can lead to poor model performance. But by far the most common problem in applied machine learning is overfitting.
* Overfitting is such a problem because the evaluation of ML on training data is different from evaluation we actually care the most about, namely how well the algorithm performs on unseen data.
* There are 2 imp techniques that you can use when evaluating ML algo to limit overfitting:—
  → use a resampling technique to estimate model accuracy.
  → hold back a validation dataset

Q14 How Random forest using ML can influence the efficiency of algorithm?

* High Accuracy
→ Random Forest models generally have high accuracy because they are an ensemble of multiple decision trees.
* Reduced Overfitting
→ Random forests are less prone to overfitting compared to individual decision trees.
* Handling Missing Values
→ Random Forests can handle missing values in the dataset w/o the need for imputation.
* Robustness
→ Random Forest is robust to outliers & noise in data. It can handle noisy data and still provide accurate predictions, reducing the need for extensive data cleaning efforts and making the overall process more efficient.

Q15 How do you check the accuracy in ML model.

We can calculate the accuracy of any model by dividing the correctly predicted problems by the total no. of predictions made.

$$\text{Accuracy} = \frac{\text{no. of correct predictions}}{\text{total no. of predictions}}$$

The above formula is very useful for calculating the accuracy of any model. It provides a simple understanding of a binary classification problem.

Q19 How Precision & Recall calculated in ML.

Precision

→ Precision measures the % age of predictions made by model that are correct.

→ $\text{Precission} = \frac{TP}{TP + FP}$

→ TP = True Positive
→ FP = False Positive

Recall

→ Recall measures the % age of relevant data points that were correctly identified by model.

→ $\text{Recall} = \frac{TP}{TP + FN}$

→ TP = True Positive
→ FN = False Negative

Q20 How CART model can be created from data?

* Creating a CART model involves selecting i/p variable & split points on those variables until a suitable tree is constructed.
* The selection of which i/p variable to use and the specific split or cut-point is chosen using a greedy algorithm to minimize a cost function.
* Tree construction ends using a predefined stopping criterion, such as min no. of training instances assigned to each leaf node of the tree.
* Tree can be constructed using following algorithm:-
  - → Greedy algorithm
  - → Stopping Criterion
  - → Tree pruning

Q21 How errors are calculated in linear Regression

→ Linear regression most often uses mean Square error (MSE) to calculate the error of the model. MSE is calculated by:-

1. measuring the distance of the observed y-values from the predicted y-values at each value of x
2. Square each of these distances
3. Calculating the mean of each of the Squared distances.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

n = no. of data points $\hat{Y}_i$ = predicted values

Y_i = observed values

Q22 Explain Tree Pruning

* Tree pruning in ML refers to the process of reducing the size of a decision tree by removing certain branches of the tree that do not provide significant information or do not improve the model's predictive accuracy
* It is used to avoid overfitting, where the tree fits the training data too closely and performs poorly on unseen or new data.
* 2 types of pruning techniques
  - → pre-pruning
  - → post-pruning
* Pruning helps in creating simpler and more interpretable models that generalize well for unseen data as it reduces the risk of overfitting and improves tree's efficiency

Q23 Explain learning methods of ML

* Batch learning

→ In many cases, we have end-to-end ML systems in which we need to train the model in one go by using whole available training data. Such kind of learning method is called batch learning.

* Online learning

→ In this method, the training data is supplied in multiple incremental batches, called mini batches to algorithm.

* Instance based learning
→ this method builds the model by doing generalisation based on i/p data.

* Model based learning
→ In the method an iterative process takes place on ML models that are built based on various model parameters called hyperparameters and in which i/p data is used to extract features.

Q29

Linear Algorithm	Non-linear algorithm
Assume a linear relationship b/w input feature & O/p predications.	Do not assume a linear relationship b/w i/p features & o/p predictio
Represent data using straight lines (in 2D) or hyperplanes (in multiple dimensions)	Can capture complex non-linear patterns & relationships in the data.
Computationally Efficient & easy to interpret, ~~suitable for simpler relationship in data~~	Computationally intensive and may require larger larger data.
Example linear Regression linear SVM.	Example decision trees neural networks

Q5 Explain Generalization in ML

* Generalization refers to how well the concepts learned by a ML model apply to specific examples not seen by the model when it was learning
* The goal of good ML model is to generalize well from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen.
* There is a terminology used in ML when we talk about how well a ML model learns & generalizes to new data, namely overfitting & underfitting.
* Overfitting & underfitting are the 2 biggest causes for poor performance of ML algo

Q6 Explain Linear Regression Model

→ Linear Regression is a linear model e.g. a model that assumes the linear relationship b/w the input variables (x) and the single o/p variable(y).

→ More specifically, that y can be calculated from a linear combination of i/p variables (x). When there is a single i/p variable (x), the method is referred to as Simple Linear regression $Y = B0 + B1(x)$

→ When there are multiple i/p variables the method is referred as multiple Linear Regression

* Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is ordinary least squares.

Q38 How ML & AI are coordinated?

* AI's big picture
→ AI is the overarching concept where machines simulate human intelligence, including problem solving & decision-making.

* ML's specialized Role
→ ML is a subset of AI. It involves algorithms learning from data, improving tasks w/o being explicitly programmed.

* Data-powered Learning
→ ML algo analyze data, identifying patterns and making predictions. This data-driven learning enhances AI's abilities.

* Continuous Improvement
→ ML enables AI systems to learn from new data, ensuring they adapt and improve, making them more intelligent & effective over time.