1. In the sense of machine learning, what is a model? What is the best way to train a model?

A machine learning model is a file that has been trained to recognize certain types of

patterns. You train a model over a set of data, providing it an algorithm that it can use to

reason over and learn from those data.

1. **Model Naming**
2. **Data Type Selection**
3. **Data Upload**
4. **Type category**
5. **Start Training**
6. In the sense of machine learning, explain the "No Free Lunch" theorem.

The “no free lunch” (NFL) theorem for supervised machine learning is a theorem that essentially implies that **no single machine learning algorithm is universally the best-performing algorithm for all problems**.

3. Describe the K-fold cross-validation mechanism in detail.

Split the dataset into k groups

For each unique group:

Take the group as a hold out or test data set

Take the remaining groups as a training data set

Fit a model on the training set and evaluate it on the test set

Retain the evaluation score and discard the model

Summarize the skill of the model using the sample of model evaluation scores

4. Describe the bootstrap sampling method. What is the aim of it?

The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples. Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen.

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

5. What is the significance of calculating the Kappa value for a classification model? Demonstrate how to measure the Kappa value of a classification model using a sample collection of results.

It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless.

values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

6. Describe the model ensemble method. In machine learning, what part does it play?

 Ensemble methods means combining multiple models to produce better results than individually trained models.

Ensemble are of 3 types.

1.Bagging

2. Boosting

3. Stacking

7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.

Descriptive modeling is a mathematical process that **describes real-world events and the relationships between factors responsible for them**. The process is used by consumer-driven organizations to help them target their marketing and advertising efforts.

Ex: Customer segmentation

8. Describe how to evaluate a linear regression model.

Evaluation of a linear regression model can be done using R-square. R square is calculated as the sum of squared errors in predictions made, divided by summation of all sum of squares. R square measures how much of the change in target variable can be explained by the linear regressor. Its value ranges from 0 to 1 where 0 means poor performance and 1 means good. Some other techniques which can be used to evaluate a linear regression model are:

1. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
2. Mean Absolute Error(MAE)

9. Distinguish :

1. Descriptive vs. predictive models

* Descriptive models are built to identify trends and underlying patterns.
* Predictive models are built to predict a dependent variable value.
* Most of descriptive models are built using unsupervised machine learning.
* Most of predictive models are built using classification and regression models.
* Example for descriptive model: Finding why consumers are engaging more with a social media post.
* Example for predictive model: Predicting the chances of cancer in a patient.

2. Underfitting vs. overfitting the model

* Underfitting is a situation arising when the hypothesis is way too simple, or when the machine learning model is way too simple to produce good results.
* Overfitting is a situation arising when the hypothesis is way too complex, or when the machine learning model is way too complex to produce good results.
* Underfitting causes a model to produce poor results due to heavily simplified algorithm reacting lightly to changes in the unseen data for independent variables from the training data.
* Overfitting makes a model produce poor results due to slightest variations in the unseen data for independent variables from the training data
* Underfitting is also called High Bias.
* Overfitting is also called High variance

3. Bootstrapping vs. cross-validation

* Boostrap sampling is a method of sampling in which the repeated sampling is done with replacement using a data D in random draws over which machine learning models are trained for better performance.
* Cross validation is a method used to check the efficacy of the machine learning model on test data.
* End goal of bootstrapping is to reduce overfitting and increase performance.
* End goal of cross validation is only to produce test scores to check efficacy of model
* Bootstrapping is best employed in Random Forest Classifier.
* Cross Validation is best employed using K-fold cross validation technique.

10. Make quick notes on:

1. LOOCV.

The Quick notes on: LOOCV or Leave One Out Cross Validation is a form of K-fold

cross validation where only one observation is left out for validation purpose while the

rest of the data is used for model training each iteration. It is computationally taxing and

should only be used for data with low dimensionality.

2. F-measurement

Harmonic mean of Precision score and recall score is called F-measurement or F-score.



3. The width of the silhouette

Estimate of average inter cluster distance to give efficacy/performance of cluster algorithms

is called width of the silhouette. It can also be defined as how identical/similar a data point 'x'

is to the data points inside the cluster to which x is assigned. Its value ranges from -1 to 1

where 1 means good and -1 means bad.

4. Receiver operating characteristic curve

Curve plotted between True Positive Rate and False Positive Rate is Receiver Operating Characteristics curve and is used to find the area under the curve for ROC-AUC score for binary classification evaluation. True Positive Rate and False Positive Rate are calculated for different thresholds values where thresholds take values starting from the highest probability scores assigned to data points and goes up to the lowest probability score. The curve is impacted by presence of outliers, and simple models. Extensions can be made to this curve to suit multiclass classification evaluation requirements.