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.

ML is **a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method**. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.

Training the model in Machine Learning

**3 steps to training a machine learning model**

Step 1: Begin with existing data. Machine learning requires us to have existing data—not the data our application will use when we run it, but data to learn from.

Step 2: Analyze data to identify patterns.

Step 3: Make predictions.

2. In the sense of machine learning, explain the &quot;No Free Lunch&quot; theorem.

The No Free Lunch theorems **prove that under a uniform distribution over induction problems (search problems or learning problems), all induction algorithms perform equally**.

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.

K-fold Cross-Validation is **when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data**. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation.

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

The bootstrap method is a resampling technique used **to estimate statistics on a population by sampling a dataset with replacement**. It can be used to estimate summary statistics such as the mean or standard deviation.

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.

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

Ensemble methods is a machine learning technique that **combines several base models in order to produce one optimal predictive model** . To better understand this definition lets take a step back into ultimate goal of machine learning and model building.

It is seen as a part of **artificial intelligence**. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

7. What is a descriptive model & main purpose? Give examples of real-world problems that

descriptive models were used to solve.

Machine learning is a subfield of artificial intelligence, which is broadly defined as **the capability of a machine to imitate intelligent human behavior**. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.

8. Describe how to evaluate a linear regression model.

The linear regression model **describes the dependent variable with a straight line that is defined by the equation Y = a + b × X, where a is the y-intersect of the line, and b is its slope.**

Evaluation methods for Regression Models depend on the ways of calculating the difference between the actual and predicted values. There are methods such as the **Sum Squared Error, Mean Squared Error and Root Mean Squared Error** which are among the various ways of calculating this difference.

**There are 3 main metrics for model evaluation in regression:**

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

9. Distinguish :

1. Descriptive vs. predictive models

A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a Predictive model, it identifies patterns found in past and transactional data to find risks and future outcomes.

1. Underfitting vs. overfitting the model

**Underfitting means that your model makes accurate, but initially incorrect predictions**. In this case, train error is large and val/test error is large too. Overfitting means that your model makes not accurate predictions. In this case, train error is very small and val/test error is large

1. Bootstrapping vs. cross-validation

In summary, Cross validation splits the available dataset to create multiple datasets, and Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. Bootstrapping it is not as strong as Cross validation when it is used for model validation.

10. Make quick notes on:

1. LOOCV

LOOCV(Leave One Out Cross-Validation) is **a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set**. In LOOCV, fitting of the model is done and predicting using one observation validation set.

1. F-measurement

An F-score is the harmonic mean of a system’s precision and recall values. It can be calculated by the following formula: 2 x [(Precision x Recall) / (Precision + Recall)]. Criticism around the use of F-score values to determine the quality of a predictive system is based on the fact that a moderately high F-score can be the result of an imbalance between precision and recall and, therefore, not tell the whole story. On the other hand, systems at a high level of accuracy struggle to improve precision or recall without negatively impacting the other.

Critical (risk) applications that value information retrieval more than accuracy (i.e., producing a large number of false positives but virtually guaranteeing that all the true positives are found) can adopt a different scoring system called F2 measure, where recall is weighed more heavily. The opposite (precision is weighed more heavily) is achieved by using the F0.5 measure.

1. The width of the silhouette

The Average Silhouette Width (ASW) of a clustering is ̄ a ( i ) is **the average distance of to points in the cluster to which it was assigned**, and is the average distance of to the points in the nearest cluster to which it was not assigned.

The silhouette value is **a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)**. The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.