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

A machine learning training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from.

There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.

Supervised learning is possible when the training data contains both the input and output values. Each set of data that has the inputs and the expected output is called a supervisory signal. The training is done based on the deviation of the processed result from the documented result when the inputs are fed into the model.

Unsupervised learning involves determining patterns in the data. Additional data is then used to fit patterns or clusters. This is also an iterative process that improves the accuracy based on the correlation to the expected patterns or clusters. There is no reference output dataset in this method.

2. In the sense of machine learning, explain the "No Free Lunch" theorem.

The No Free Lunch Theorem, often abbreviated as NFL or NFLT, is a theoretical finding that suggests all optimization algorithms perform equally well when their performance is averaged over all possible objective functions.

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

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.

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.

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 . The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

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. These systems are used everywhere by search engines, e-commerce websites (Amazon), entertainment platforms (Google Play, Netflix), and multiple web & mobile apps.

8. Describe how to evaluate a linear regression model.

* R Square/Adjusted R Square
* Mean Square Error(MSE)/Root Mean Square Error(RMSE)
* Mean Absolute Error(MAE)

9. Distinguish :

1. Descriptive vs. predictive models

Descriptive Analytics will give you a vision into the past and tells you: what has happened? Whereas the Predictive Analytics will recognize the future and tells you: What might happen in future?

Descriptive Analytics uses Data Aggregation and Data Mining techniques to give you knowledge about past but Predictive Analytics uses Statistical analysis and Forecast techniques to know the future.

Descriptive Analytics is used when you need to analyze and explain different aspects of your organization whereas Predictive Analytics is used when you need to know anything about the future and fill the information that you do not know.

2. Underfitting vs. overfitting the model

Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the 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 low bias and high variance.

Underfitting occurs when our machine learning 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 early 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. In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

3. Bootstrapping vs. cross-validation

Bootstrapping is any test or metric that relies on random sampling with replacement.It is a method that helps in many situations like validation of a predictive model performance, ensemble methods, estimation of bias and variance of the parameter of a model etc. It works by performing sampling with replacement from the original dataset, and at the same time assuming that the data points that have not been choses are the test dataset.

Cross validation is a procedure for validating a model's performance, and it is done by splitting the training data into k parts. We assume that the k-1 parts is the training set and use the other part is our test set. We can repeat that k times differently holding out a different part of the data every time. Finally, we take the average of the k scores as our performance estimation. Cross validation can suffer from bias or variance. Increasing the number of splits, the variance will increase too and the bias will decrease. On the other hand, if we decrease the number of splits, the bias will increase and the variance will decrease.

10. Make quick notes on:

1. LOOCV.

It 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.

2. F-measurement

F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

3. The width of the silhouette

The Average Silhouette Width (ASW) is a popular cluster validation index to estimate the number of clusters.

4. Receiver operating characteristic curve

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.