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

Answer: If we completely look from machine learning model is nothing but a complex mathematical equation which is self-taught by computers based on the past experiences and the various set of data fed to the ever changing mathematical functions. The best way to train a model is by feeding that complex mathematical equation as much data we can with max variances so that it will be handle the extreme unknown cases without any difficulty.

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

Answer: No free lunch theorem states that all of the optimization algorithms will perform almost under some specific constraints and there is no single best optimization algorithm or machine learning algorithm.

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

Answer: K fold cross validation is method to check the skill of model on new data. For performing the k fold cross validation, we do the following steps:

1. Shuffle the dataset randomly
2. Split entire dataset into k subgroups
3. Take k-1 random groups as training sets and 1 remainder set as testing set from an entire set of available training datasets.
4. Train the model on training datasets and evaluated on testing dataset. This same process is repeated k times. We retain evaluation score and discard the model for every iteration.
5. We now finally 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?

Answer: Following are the steps to be done in a bootstrap sampling methods:

1. Choose the number of bootstrap samples to perform.
2. Choose a sample size.
3. For each bootstrap sample
4. Draw a sample with replacement with chosen size
5. Calculated the stats on the sample
6. Calculate the mean of the stats.

There is a possibility that a sample selected in one bootstrap sample maybe repeated in multiple bootstrap samples.

The main aim of bootstrap sampling method is to predict if there is any unexpected uncertainity in the dataset that we have on our end.

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.

Answer: Kappa is a statistical measure of how closely the instances classified by the machine learning classifier matched the data labelled as ground truth and thereby controlling the accuracy of the classifier model by expected accuracy.

K = (accuracy – random accuracy)/(1-random accuracy)

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

Answer: Ensemble modelling is a method of getting the overall prediction by using majority vote concept to get final result when all the models have produced output on their end. It helps in getting overall idea of prediction for an unknown dataset passed through the model.

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

Answer: A descriptive machine learning model’s main aim is to summarize, classify and extract rules to answer to what happens has happened in the past.

8. Describe how to evaluate a linear regression model.

Answer: Linear regression is based on the difference /error calculated between the original value and predicted value from model such that values predicted from regression model on an unknown test data is the best one with errors close to 0.

9. Distinguish :

1. Descriptive vs. predictive models

Answer: A descriptive model will simply use the past data stored in any location at your end will try to draw conclusions and describe the events happened in the past.

Predictive models will find the pattern in past and current data and will try to predict the future risks and outcomes.

2. Underfitting vs. overfitting the model

Answer: Underfitting is a condition when your ML model performs very bad on training data and test data based on performance metrics we’ve decided to test the model.

Overfitting is a situation where your model performs really well but it’s performance goes down when an unknown data is passed through the same model.

3. Bootstrapping vs. cross-validation

Answer: Bootstrapping is the dividing/sampling of data at the training phase itself so that each divided data goes to different models for training purposes.

Cross validation is mostly done for the testing part of the model by dividing some small chunk of training data into a sample test set such that we are able to train and test our model for every possible unknown dataset.

10. Make quick notes on:

1. LOOCV.

Answer: It is a small variant of k fold cross validation where only point is left for testing and rest of the points are used for training process. This process is repeated n times where n is total # of data points.

2. F-measurement

Answer: F1 score of F measurement = (2\*precision\*recall)(precision +recall)

3. The width of the silhouette

Answer: Silhoutte width ranges from -1 to 1.

4. Receiver operating characteristic curve

ROC curve is graphical representation between false +ve rate and true +ve rate to indicate which models better than the decided threshold and in most of the cases it is 50% where we try to increase the true +ve rate as much as we can try to reduce the false +ve rate.