Machine learning 06

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

Ans:- 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.

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

Ans:- 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.

Ans:- 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?

Ans:- 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.

Ans:- 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.

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

Ans:- 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.

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

Ans:- A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it. A geometric model or spatial model is a descriptive model that represents geometric and/or spatial relationships.

8. Describe how to evaluate a linear regression model.

1. Ans:- 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.

2. Underfitting vs. overfitting the model :- 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.

3. 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

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.

2. F-measurement :- The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

3. The width of the silhouette :- 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.

4. Receiver operating characteristic curve :- An ROC curve is a plot of sensitivity on the y axis against (1−specificity) on the x axis for varying values of the threshold t. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. The ROC curve for the gold standard is the line connecting (0,0) to (0,1) and (0,1) to (1,1).