MACHINE LEARNING

1. Which of the following in sk-learn library is used for hyper parameter tuning?

Ans. A) GridSearchCV() B) RandomizedCV()

2. In which of the below ensemble techniques trees are trained in parallel?

Ans. A) Random forest

3. In machine learning, if in the below line of code: sklearn.svm.SVC (C=1.0, kernel='rbf', degree=3) we increasing the C hyper parameter, what will happen?

Ans. B) The regularization will decrease

4. Check the below line of code and answer the following questions: sklearn.tree.**DecisionTreeClassifier**(*criterion='gini',splitter='best',max_depth=None, min_samples_split=2) Which of the following is true regarding max_depth hyper parameter?

Ans. A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.

5. Which of the following is true regarding Random Forests?

Ans. C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.

6. What can be the disadvantage if the learning rate is very high in gradient descent?

Ans. A) Gradient Descent algorithm can diverge from the optimal solution.

7. As the model complexity increases, what will happen?

Ans. B) Bias will decrease, Variance increase

8. Suppose I have a linear regression model which is performing as follows: Train accuracy=0.95 and Test accuracy=0.75 Which of the following is true regarding the model?

Ans. A) model is underfitting

Q9 to Q15 are subjective answer type questions, Answer them briefly.

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index of the dataset.

Ans. P(Past Trend=Positive): 60/100

P(Past Trend=Negative): 40/100

If (Past Trend = Positive & Return = Up), probability = 4/6
If (Past Trend = Positive & Return = Down), probability = 2/6
Gini Index = 1 - ((4/6)^2 + (2/6)^2) = 0.45

If (Past Trend = Negative & Return = Up), probability = 0
If (Past Trend = Negative & Return = Down), probability = 4/4
Gini Index = 1 - ((0)^2 + (4/4)^2) = 0

Weighted sum of the Gini Indices can be calculated as follows:
Gini Index for Past Trend = (60/100)0.45 + (40/100)0 = 0.27

10. What are the advantages of Random Forests over Decision Tree?

Ans. Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Ans. Scaling is required to rescale the data and it's used when we want features to be compared on the same scale for our algorithm. And, when all features are in the same scale, it also helps algorithms to understand the relative relationship better.

The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.

Ans. Having features on a similar scale will help the gradient descent converge more quickly towards the minima. Specifically, in the case of Neural Networks Algorithms, feature scaling benefits optimization by: It makes the training faster. It prevents the optimization from getting stuck in local optima.

13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

Ans. Accuracy is not a good metric for imbalanced datasets.This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

14. What is “f-score" metric? Write its mathematical formula.

Ans. 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'.

Formula:-

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

15. What is the difference between fit(), transform() and fit_transform()?

Ans. The fit() method helps in fitting the data into a model.

transform() method helps in transforming the data into a form that is more suitable for the model.

Fit_transform() method, on the other hand, combines the functionalities of both fit() and transform() methods in one step.