MACHINE LEARNING

1. Which of the following in sk-learn library is used for hyper parameter tuning?
   A) GridSearchCV()   B) RandomizedCV()
   C) K-fold Cross Validation   D) All of the above
   Ans: D

2. In which of the below ensemble techniques trees are trained in parallel?
   A) Random forest   B) Adaboost
   C) Gradient Boosting   D) All of the above
   Ans: A

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?
   A) The regularization will increase   B) The regularization will decrease
   C) No effect on regularization   D) kernel will be changed to linear
   Ans: A

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?
   A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.
   B) It denotes the number of children a node can have.
   C) both A & B
   D) None of the above
   Ans: C

5. Which of the following is true regarding Random Forests?
   A) It's an ensemble of weak learners.
   B) The component trees are trained in series
   C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.
   D)None of the above
   Ans: A

6. What can be the disadvantage if the learning rate is very high in gradient descent?
   A) Gradient Descent algorithm can diverge from the optimal solution.
   B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.
   C) Both of them
   D) None of them
   Ans: D

7. As the model complexity increases, what will happen?
   A) Bias will increase, Variance decrease   B) Bias will decrease, Variance increase
   C) both bias and variance increase   D) Both bias and variance decrease.
   Ans: B

MACHINE LEARNING

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?
   A) model is underfitting B) model is overfitting
   C) model is performing good D) None of the above
   Ans: A

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 and entropy of the dataset.

Ans:

Gini Index = subtracting the sum of squared probabilities of each class from one.
= 1- (0.4^2 + 0.6^2) = 0.48
Entropy = - (0.4 log 0.4 + 0.6 log 0.6) = 0.29

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: Feature Scaling is a method to transform the numeric features in a dataset to a standard range so that the performance of the machine learning algorithm improves. It can be achieved by normalizing or standardizing the data values. 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: Feature scaling is a method to unify self-variables or feature ranges in data. In data processing, it is usually used in data pre-processing. Because in the original data, the range of variables is very different. Feature scaling is a necessary step in the calculation of stochastic gradient descent.

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 does not holds good for imbalanced data. In business scenarios, most data won't be balanced and so accuracy becomes poor measure of evaluation for our classification model.

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

Ans: In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall. 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)].

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.

NEHA ACHARYA (Batch: 35)