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

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

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

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

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

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

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.

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

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 and entropy of the dataset.
10. What are the advantages of Random Forests over Decision Tree?
11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

MACHINE LEARNING

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
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?
14. What is “f-score" metric? Write its mathematical formula.
15. What is the difference between fit(), transform() and fit_transform()?

9: Gini index is calculated by subtracting the sum of squared probabilities of each class from one.

10: Random Forest is simply a collection of decision trees whose results are aggregated into one final result. It reduces overfitting in decision trees and helps to improve the accuracy. It is flexible to both classification and regression problems. It works well with both categorical and continues values. It automates the missing values in the dataset.,

11: Scaling can make difference between a weak machine learning model and a better one.The most common techniques of feature scaling are Normalization and standardization.

12: Advantages of gradient descent algorithms are its computational efficient. It produces a stable error gradient and a stable convergence.

13: The most common metric used to evaluate the performance of a classification predictive model is classification accuracy. For an imbalanced dataset accuracy is no longer a proper measure, since it doesn’t distinguish between numbers of correctly classified examples of different classes. Hence it may lead to erroneous conclusion.

14: F-score is a measure of test’s accuracy. It is calculated from precision and recall of the test.F1 score is the harmonic mean of precision and recall.

15 : fit() calculates the values of parameters.transform function applies the value of the parameters ontheactualdata andgivesthenormalizedvalue.Thefittransform()functionperformsbothinthesame step. Same value is got if we perform both steps or in a single step.