Machine Learning Assignment_7

QUES 1: Which of the following in SK-Learn library is used for hyper parameter tuning?

Answer. A) GridSearchCV()

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

Answer. A) Random Forest.

QUES 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?

Answer. D) Kernel will be changed to linear.

QUES 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?

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

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

Answer. A) It's an ensemble of weak learners.

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

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

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

Answer. B) Bias will decrease, Variance increase.

QUES 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?

Answer. B) Model is Overfitting.

QUES 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.

Answer. Gini index = 1- (p(A)2 + p(B)2) = 1-((0.4)2+(0.6)2) = 0.48

Entropy = -(p(A) log2(p(A)) + p(B) log2(p(B))) = -(0.4*log2(0.4) + 0.6*log2(0.6))) = 0.97.

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

Answer. The advantages of Random Forests over Decision Tree is:

1. Random forest algorithm avoids and prevents overfitting by using multiple trees.
2. Improved accuracy.

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

Answer. The need of scaling all numerical features in a dataset is to check that the performance of the machine learning algorithm improves.

Any two techniques used for scaling are : 1) Z-Scaling 2) Min-Max Scaling.

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

Answer. We can use fixed learning rate during training without worrying about learning rate decay.

It has straight trajectory towards the minimum and it is guaranteed to converge in theory to the global minimum if the loss function is convex and to a local minimum if the loss function is not convex.

QUES 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?

Answer. 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.

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

Answer. 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)].

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

Answer. fit(): The fit() method will allow us to get the parameters of the scaling function.

transform():The transform() method will transform the dataset to proceed with further data analysis steps.

fit_transform(): The fit_transform() method will determine the parameters and transform the dataset.