MACHINE LEARNING

ASSIGNMENT - 7

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

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- 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?
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- 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
ANS- C) Both 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.
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?
B) Bias will decrease, Variance increase B) model is overfitting
C) model is performing good D) None of the above
ANS- B) Bias will decrease, Variance increase

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.

ANS-

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 used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
The most common techniques of feature scaling are Normalization and Standardization

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
ANS- The main advantages:

- • 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.

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- 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(data) method is used to compute the mean and std dev for a given feature to be used further for scaling. The transform(data) method is used to perform scaling using mean and std dev calculated using the . fit() method. The fit_transform() method does both fits and transform