MACHINE LEARNING

ASSIGNMENT - 7

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

A) GridSearchCV()

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

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

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?

B) It denotes the number of children a node can have.

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

B) The component trees are trained in series

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

D) None of them

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

A) Bias will increase, 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?

B) model is overfitting

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?

Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate.This gives accurate and precise results.

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

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.

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?

accuracy does not holds good for imbalanced data. 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.

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’.

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

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

- 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.