MACHINE LEARNING

ASSIGNMENT - 7

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

D) All of the above

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

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

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.

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

D)None of the above

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

C) Both of them

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

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?

C) model is performing good

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.

A9. Gini Index: (0.4)*(0.4)+(0.6)*(0.6)= 0.16+0.36 = 0.52

Entropy: -0.4 log(0.4)- 0.6log(-0.6)= 0.970

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

A10. Random forests give better performance than decision trees as it does everything for reducing the number of variables , treating missing values , outliers values and exploring data. Random forests perform better than bagged trees as it de-correlates the trees.

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

A11. Feature scaling is essential for machine learning algorithms, if not scaled the feature with a higher value range will start dominating the dataset. ML algorithm works better when features are relatively on a similar scale also algorithms converge faster when features are relatively smaller or closer to normal distribution.

Scaling techniques:

- StandardScaler()
- Normalizer()

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

A12. Gradient decent will reach the minimum cost faster if the shape of the cost function is not skewed or distorted. This can be achieved by rescaling input variables to same range.

Hence scaling helps in faster optimisation of gradient decent.

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?

A13. In the case of imbalanced dataset accuracy metric can be misguiding, as high metrics doesn't show prediction capacity for the minority class.

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

A14. The F-score (also known as the F1 score or F-measure) is a metric used to evaluate the performance of a Machine Learning model. It combines precision and recall into a single score.

F-measure formula:

- F-score = 2 * (precision * recall) / (precision + recall)

The F-score ranges from 0 to 1, with higher values indicating better performance.

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

- fit() : used for generating learning model parameters from training data
- transform() : parameters generated from fit() method,applied upon model to generate transformed data set.
- fit_transform() : combination of fit() and transform() api on same data set