ASSIGNMENT – 7

MACHINE LEARNING

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

Answer: D) All of the above

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

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?

Ans: 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?

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?

Ans: A) It's an ensemble of weak learners.

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

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

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

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?

Ans: A) model is underfitting

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: To calculate the Gini index and entropy of the dataset, we first need to calculate the probabilities of each class:

Probability of class A = 0.4

Probability of class B = 0.6

Gini index:

Gini index = 1 - (probability of class A)^2 - (probability of class B)^2

Gini index = 1 - (0.4)^2 - (0.6)^2

Gini index = 0.48

Entropy:

Entropy = - (probability of class A) * log2(probability of class A) - (probability of class B) * log2(probability of class B)

Entropy = - 0.4 * log2(0.4) - 0.6 * log2(0.6)

Entropy = 0.97

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

Ans: Some advantages of Random Forests over Decision Trees are:

1. **Random Forests reduce overfitting and increase the accuracy of the model by combining multiple decision trees.**
2. **Random Forests can handle missing data and maintain accuracy even when a large proportion of the data is missing.**
3. **Random Forests can handle large datasets with high dimensionality and many features, without requiring feature selection or feature engineering.**
4. **Random Forests can provide measures of feature importance, which can help in feature selection and understanding the important features in the dataset.**

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

Ans: Scaling all numerical features in a dataset is necessary because the scales of the features may vary widely, and some features may dominate the others in the calculation of distances or similarities between data points. Scaling the features brings them to a similar scale and avoids the dominance of any particular feature. Two commonly used techniques for scaling are:

1. **Min-Max scaling: This technique scales the features to a range between 0 and 1, by subtracting the minimum value of the feature and dividing by the difference between the maximum and minimum values.**
2. **Standardization: This technique scales the features to have zero mean and unit variance, by subtracting the mean value of the feature and dividing by the standard deviation.**

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

Ans: Scaling provides the following advantages in optimization using gradient descent algorithm:

1. **Scaling helps gradient descent algorithm converge faster, as it reduces the number of iterations required to reach the minimum.**
2. **Scaling helps gradient descent algorithm avoid getting stuck in local minima or saddle points, by allowing it to move smoothly and consistently towards the global minimum.**
3. **Scaling helps gradient descent algorithm avoid overshooting the minimum or oscillating around it, by providing a smoother gradient and avoiding large steps.**

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: In case of a highly imbalanced dataset for a classification problem, accuracy is not a good metric to measure the performance of the model. This is because accuracy can be misleading in such cases, as a model that simply predicts the majority class all the time can have a high accuracy even though it is not useful. Instead, metrics like precision, recall, F1-score, or area under the ROC curve (AUC-ROC) are more appropriate for measuring the performance of the model in such cases.

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

Ans: F-score (also known as F1-score) is a metric that combines precision and recall of a classification model into a single score. It is a harmonic mean of precision and recall, and ranges between 0 and 1, where 1 is the best score. The mathematical formula for F-score is:

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

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

Ans: In scikit-learn library, fit(), transform() and fit_transform() are methods used in preprocessing and feature engineering of the data:

1. **fit() method is used to fit the estimator on the data, and learn the parameters from the data. For example, in a scaler, it learns the mean and standard deviation of the data.**
2. **transform() method is used to transform the data based on the parameters learned in the fit() method. For example, in a scaler, it scales the data using the mean and standard deviation learned in the fit() method.**
3. **fit_transform() method is used to perform both the fit() and transform() methods in a single step. It first fits the estimator on the data and learns the parameters, and then transforms the data based on the learned parameters.**