MACHINE LEARNING

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

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

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

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

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

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

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.

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?
   A) model is underfitting
   B) model is overfitting
   C) model is performing good
   D) None of the above

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

MACHINE LEARNING

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?
14. What is “f-score" metric? Write its mathematical formula.
15. What is the difference between fit(), transform() and fit_transform()?

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% andpercentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

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

P(A) = 0.4
P(B) = 0.6
Gini index:

Gini = 1 - (P(A)^2 + P(B)^2) = 1 - (0.4^2 + 0.6^2) = 0.48

Entropy:

Entropy = -P(A) * log2(P(A)) - P(B) * log2(P(B))
= -(0.4 * log2(0.4)) - (0.6 * log2(0.6))
= 0.9710

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

Random Forests have several advantages over Decision Trees, including:

Reducing overfitting: Random Forests use a technique called bagging, which reduces the variance and overfitting that is common in Decision Trees. By building multiple trees on different subsets of the data and combining their results, Random Forests can provide more accurate and stable predictions.

Handling missing values: Random Forests can handle missing values in the data without any preprocessing, unlike Decision Trees which require imputation or deletion of missing values.

Providing feature importance: Random Forests can provide a measure of feature importance, which can help in feature selection and identifying the most relevant variables for the target variable.

Handling large datasets: Random Forests can handle large datasets with high dimensionality, unlike Decision Trees which can become computationally expensive and may overfit in such cases.

Providing robustness to outliers: Random Forests are less sensitive to outliers and noise in the data than Decision Trees, as the random selection of features and data points during the training process helps to mitigate the impact of outliers.

MACHINE LEARNING

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

Scaling numerical features is important because many machine learning algorithms assume that all features are on the same scale and have the same variance. If the features are not on the same scale, some features may dominate others in the modeling process, leading to biased and inaccurate results. Scaling ensures that all features contribute equally to the model.

Two techniques used for scaling are:

Min-Max Scaling: This technique scales the features to a fixed range (e.g., 0 to 1) by subtracting the minimum value and dividing by the range of the feature.

Standardization: This technique scales the features to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation of the feature.

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

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

Convergence speed: Scaling ensures that the optimization algorithm converges faster to the minimum value of the cost function.

Avoiding oscillations: Scaling ensures that the optimization algorithm avoids oscillations or zigzagging during the convergence process.

Accurate computation: Scaling ensures that the optimization algorithm performs accurate computations of the gradients, which improves the quality of the solution.

Better conditioning: Scaling ensures that the optimization problem is well-conditioned, which means that the gradients of the cost function are well-scaled and well-behaved.

Two commonly used scaling techniques are:

Standardization: This technique scales the data so that it has zero mean and unit variance.

Normalization: This technique scales the data so that it lies in the range [0,1] or [-1,1].

13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric tomeasure the performance of the model. If not, why?

In case of a highly imbalanced dataset for a classification problem, accuracy may not be a good metric to measure the performance of the model. This is because accuracy does not take into account the class imbalance and can be misleading. For instance, if a model is trained on a dataset where one class has a majority of the samples and the other class has very few samples, then the model may predict the majority class more accurately, leading to a high overall accuracy. However, this may not reflect the true performance of the model, especially for the minority class.

In such scenarios, metrics like precision, recall, F1-score, and area under the **ROC curve (AUC-ROC)** can be more informative as they take into account the true positive, false positive, true negative, and

MACHINE LEARNING

false negative rates for each class.

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

The F-score is a metric used to evaluate the performance of a classification model. It is the harmonic mean of precision and recall.

The formula for F-score is:

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

where precision = true positives / (true positives + false positives)

and recall = true positives / (true positives + false negatives)

The F-score takes into account both precision and recall and provides a single score that balances the trade-off between them. It is useful when the classes are imbalanced and the accuracy metric may not be suitable for evaluating the model's performance.

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

In machine learning, the **‘fit()’** method is used to train the model on a given dataset. It estimates the model parameters using the training dataset.

The **‘transform()’** method is used to apply the learned transformations to a new dataset. For example, if the model was trained on a dataset with 100 samples and 10 features, then the ‘**transform()’** method can be used to apply the same transformations on a new dataset with 50 samples and 10 features.

The ‘fit_transform()’ method is a combination of both **‘fit()’** and **‘transform()’**. It first fits the model on the training dataset and then applies the same transformations on the training dataset. This is typically used in cases where the original dataset needs to be transformed before being used for training.

In summary,**’ fit()’** is used for training the model, **‘transform()’** is used for applying transformations, and **‘fit_transform()’** is used for both training and applying transformations.