MACHINE LEARNING

Q1 to Q8, Choose the correct option:

1. **Which of the following in scikit-learn library is used for hyper parameter tuning?**
   a. GridSearchCV
   b. RandomizedCV
   c. K-fold Cross Validation
   d. All of the above

   Answer: a. GridSearchCV **and** b. RandomizedCV

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

   Answer: 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

   Answer: 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
b. It denotes the number of children a node can have
c. Both A & B
d. None of the above

Answer: 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

   Answer: a. It's an ensemble of weak learners **and**
   c. In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees

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

   Answer: 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

Answer: b. Bias will decrease, Variance increases

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

   Answer: 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?**

   Answer: To calculate the Gini index of the dataset, we use the formula:

$$Gini = 1 - (p1^2 + p2^2)$$

   Where;

   - p1 is the probability of class A
   - p2 is the probability of class B.

   In this case, p1 = 0.4 and p2 = 0.6, so,

$$Gini = 1 - (0.4^2 + 0.6^2) = 0.48$$

   To calculate the entropy of the dataset, we use the formula

$$entropy = -p1log(p1) - p2log(p2)$$

   In this case, p1 = 0.4 and p2 = 0.6, so,

$$entropy = -0.4log(0.4) - 0.6log(0.6) = 0.97$$

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

Answer: Some advantages of Random Forests over Decision Tree are:

- Random forests are less prone to overfitting than a single decision tree.
- Random forests can handle missing values and maintain accuracy.
- Random forests can be used for both classification and regression problems.
- Random forests can provide feature importance, which can be used for feature selection.

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

Answer: Scaling all numerical features in a dataset is important because many machine learning algorithms, such as those based on distance measures, are sensitive to the scale of the features. Without scaling, these algorithms would be affected by the presence of large scale features and small scale features. Two techniques used for scaling are Min-Max Scaling and Standardization.

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

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

- It helps to converge faster by reducing the oscillations in the optimization path.
- It helps to find the global minimum by reducing the chances of getting stuck in a local minimum.

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?

Answer: 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 does not take into account the imbalance of the classes and can be misleading. Other metrics such as precision, recall, F1-score, and AUC-ROC are more suitable for such cases.

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

Answer: F-score is a metric that balances precision and recall and is commonly used in classification problems. 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()?

Answer: fit() is used to train a model on the training data, transform() is used to apply a pre-trained model to the new data and fit_transform() is used to train a model on the training data and then apply it to the new data.

The difference between fit() and transform() is that fit() learns the parameters of the model, i.e., it trains the model on the training

data, while transform() applies the already trained model to the new data. On the other hand, fit_transform() combines the functionality of fit() and transform() in one step, it first trains the model on the training data and then applies it to the new data in one step.