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

Correct option: A, B

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

Correct option: A

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

Correct option: C

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

Correct option: D

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

Correct option: C

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

Correct option: B

MACHINE LEARNING

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.

Correct option: A

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

Correct option: B

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.

Entropy = - (log2 (.60)*0.60 + log2 (.40)*0.40) = 0.97

Gini index= 1-(0.40**2 + 0.60**2) = 0.48

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

Advantages of Random Forests over Decision Tree includes:

- Reduction in over fitting: by averaging several trees, there is a significantly lower risk of over fitting.
- Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.

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

It is necessary to scale the features such that each of the features are in the same scale for instance -1 to 1 or 0 to 1, otherwise the result may be skewed towards a particular feature which you may not want.

Techniques used for scaling:

- Min-Max scalar
- Normalization

MACHINE LEARNING

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

Scaling helps in causing Gradient Descent to converge much faster as standardizing all the variables on to the same scale.

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?

No, accuracy is not a good metric when it comes for unbalanced data. In unbalanced you might see the high accuracy but it will only classify one type of class best and another worst and if another class is also more important to classify then we might end up giving best result if minority class is more imp to identify.

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’

F1 = 2 * (precision * recall) / (precision + recall)

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

Fit () computes the mean and std to be used for later scaling.

Transform () uses a previously computed mean and std to auto scale the data (subtract mean from all values and then divide it by std).

Fit transform () does both at the same time.