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?
random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit overfitting as well as error due to bias and therefore yield useful result.

MACHINE LEARNING

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

    transforming your data so that it fits within a specific scale,

    min max scaler and standard scaler

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

    To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

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?

    Accuracy **is not a good metric for imbalanced datasets**. Say we have an imbalanced dataset and a badly performing model which always predicts for the majority class. This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

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

    An F-score is the harmonic mean of a system’s precision and recall values. It can be calculated by the following formula: 2 x [(Precision x Recall) / (Precision + Recall)].

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

    In the fit() method, we apply the necessary formula to the feature of the input data we want to change and compute the result before fitting the result to the transformer

    To change the data, we most likely use the transform() function, where we perform the calculations from fit() to each value in feature F

    The training data is scaled, and its scaling parameters are determined by applying a fit_transform() to the training data. The model we created, in this case, will discover the mean and variance of the characteristics in the training set.

MACHINE LEARNING

16.