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

Ans : 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

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?
   A) The regularization will increase B) The regularization will decrease
   C) No effect on regularization D) kernel will be changed to linear

Ans : The regularization will decrease

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

Ans : both A & B

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

Ans : 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

Ans : 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.

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?
   A) model is underfitting B) model is overfitting

C) model is performing good D) None of the above

Ans : 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.

Ans : The Gini Index of given dataset is 0.28 & Entropy is 0.97

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

Ans:

- Works very good with non linear data.
- Run efficiently with large dataset
- We can get good accuracy than Decision tree
- Works well with categorical and continuous data as well
- Reduce over fitting and helps to improve accuracy

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

Ans :

scaling is important to minimize impact of big range of data in dataset. If we do not scale our data then big range of number will dominant ML models which will give us falls accuracy. Scale arrange data in specific scale so that ML models can extract correct learning from it.

MACHINE LEARNING

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

Ans : It gives a better error surface shape.
It makes training faster. It prevents the optimization from getting stuck in local optima.
It is also important to apply feature scalling.

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 highly imbalanced dataset for a classification problem accuracy is not at all good matrix to measure a performance of the model achieving 90 percent classification accuracy, or even 99 percent classification accuracy, may be trivial on an imbalanced classification model.

This means that intuitions for classification accuracy developed on balance class distribution will be applied and will be wrong.

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

Ans.

Precision is a metric that calculates the percentage of correct predictions for the positive class. Recall calculates the percentage of correct predictions for the positive class out of all positive predictions that could be made. Maximizing precision will minimize the false-positive errors, whereas maximizing recall will minimize the false-negative errors.

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

Ans.

fit() :

In the fit() method, where we use the required formula and perform the calculation on the feature values of input data and fit this calculation to the transformer. For applying the fit() method (fit transform in python) we have to use .fit() in front of the transformer object.

transform() :

For changing the data we probably do transform, in the transform() method, where we apply the calculations that we have calculated in fit() to every data point in feature F. We have to use .transform() in front of a fit object because we transform the fit calculations.

it_transform()

This fit_transform() method is basically the combination of fit method and transform method, it is equivalent to fit().transform(). This method performs fit and transform on the input data at a single time and converts the data points. If we use fit and transform separate when we need both then it will decrease the efficiency of the model so we use fit_transform() which will do both the work.

FLIP ROBO