**MACHINE LEARNING**

Q1. 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

Q2. 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. D) all of the above

Q3. 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 B) The regularization will decrease

Q4. 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 c) both a and b.

Q5. 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. C) ) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees

Q6. 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

Q7. 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 and variance will increase

Q8. 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. C) model is performing good**

**Q9 to Q15 are subjective answer type questions, Answer them briefly.**

Q9. 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 gini index is 0.4 and entropy is 0.2

Q10. What are the advantages of Random Forests over Decision Tree?

Ans. Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

Q11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Ans. Feature Scaling is a method to transform the numeric features in a dataset to a standard range **so that the performance of the machine learning algorithm improves** It can be achieved by normalizing or standardizing the data values.

The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

Q12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.

Ans. **The main advantages:**

* A) We can use fixed learning rate during training without worrying about learning rate decay.
* B) It has straight trajectory towards the minimum and it is guaranteed to converge in theory to the global minimum if the loss function is convex and to a local minimum if the loss function is not convex.

Q13. 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. **Accuracy is not a good metric for imbalanced datasets**. 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

Q14. What is “f-score" metric? Write its mathematical formula.

Ans. The F-score,is a **metric used to evaluate the performance of a Machine Learning model**. It combines [**precision**](https://deepchecks.com/glossary/precision-in-machine-learning/) and [**recall**](https://deepchecks.com/glossary/recall-in-machine-learning/) into a single score.

F-measure formula:

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

Q15. 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 frontof 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.

#### fit\_transform() or fit transform sklearn

The fit\_transform() method is basically the combination of the fit method and the transform method. This method simultaneously performs fit and transform operations on the input data and converts the data points.Using fit and transform separately when we need them both decreases the efficiency of the model. Instead, fit\_transform() is used to get both works done.

v