**MACHINE LEARNING – WORKSHEET 4**

1. Which of the following in sklearn library is used for hyper parameter tuning?

Answer: A) GridSearchCV() B) RandomizedCV()

2. In which of the below ensemble techniques trees are trained in parallel?

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?

Answer: B) 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?

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?

Answer: A) It's an ensemble of weak learners

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?

Answer: C) Both of them

7. As the model complexity increases, what will happen?

Answer: C)both bias and variance increase

8. Suppose I have a linear regression model which is performing as follows:

Train accuracy=0.95

Test accuracy=0.75

Which of the following is true regarding the model?

Answer: B) model is overfitting

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:

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

Answer: Random forest is the ensemble technique, it uses multiple classification or regression algorithm to evaluate the final result, whereas the Decision Tree runs only single algorithm to evaluate the result. Random forest can have multiple Decision tree algorithm running in parallel, it uses the soft voting mechanism to come with final result. The voting mechanism counts for each prediction of algorithm that are present in random forest. At the end, it count the prediction for each class, on the basis of the prediction is made, it takes the mode of the class labels predicted by the component trees.

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

Answer: In machine learning, system only understand numerical values, if there is a vast difference between the numbers say the number ranging in thousands and tens so while performing the calculation machine learning will give more priority to the greater number. For example, suppose we have a feature weight and price wherein on value of weight is 100gram whereas one value of price is 10. So in this case algorithm will consider that 100 value is greater than 10 and it will superiority the weight feature over price. As per human there is no comparison between weight and price value, but machine will not understand. Hence to scale down all value to same scaling technique get use. It has some more advantages as well, it helps to fast computation of algorithm, for technique where algorithm is calculating difference between two points e.g. Euclidean distance or Manhattan distance it is fast. Scaling techniques are

* MinMaxScaler
* StandardScaler
* Max abs Scaler

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

Answer: As scaling helps to normalize the data range from big range to small range, it gives the advantage while computing the distances between two points or values. The scaling helps to speed up gradient descent, because θ descends quickly on small ranges and slowly on large range, and oscillates inefficiently down to the optimum when the variables are very uneven.

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: No, in case of imbalanced dataset considering accuracy score to evaluate our model performance will lead to a bad model. When the dataset is imbalanced, model usually being biased to the values that are present more. For example, if dataset is having 100 Positive values and 900 negative values, data will get biased while predicting because of high occurrence of negative values. By considering accuracy score for evaluating model performance for imbalance data it will always give the high accuracy as data trained on high occurrence more. In condition of imbalanced data we usually go for precision or recall.

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

Answer: f-score is the metric to measure the model performance, we consider f-score value when data is mostly imbalanced it considers both precision and recall, it is the harmonic mean of the precision and recall. It is best if there is some sort of balance between precision and recall in the system. Oppositely F1 score is not so high if one measure is improved at the expense of the other.

**F1 score = 2\*(Recall\*Precision)/(Recall+Precision)**

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

Answer: fit() is used for calculating the initial filling of parameters on the training data and saves them as an internal objects state.

Transform(): Use the calculated value of fit() and return modified training data

fit\_transform(): It joins the fit() and transform(). First it calls the fit() and then transform the data using transform().