WORKSHEET

MACHINE LEARNING – WORKSHEET 4

In Q1 to Q8, only one option is correct, Choose the correct option:

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

A) GridSearchCV()

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

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

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

5. Which of the following is true regarding Random Forests?

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?

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

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

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?

B) 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.

Gini index: 0.48

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

The decision tree algorithm is quite easy to understand and interpret. But often, a single tree is not sufficient for producing effective results. Random Forest is an ensemble technique which is a tree-based machine learning algorithm that leverages the power of multiple decision trees for making decisions.

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

Scaling the data makes contours smaller which means it would take shorter time to converge.

Scaling techniques: Min-Max scaling, Statandardisation

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

Without scaling the contours would be narrower and taller which means it would take longer time to converge.

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?

It is not. As in this case, accuracy has major contribution from majority of category.

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

The F1 score is the harmonic mean of the precision and recall.

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. (just a computation), nothing is given to you. "transform" uses a previously computed mean and std to autoscale the data (subtract mean from all values and then divide it by std). "fit\_transform" does both at the same time.