MACHINE LEARNING – WORKSHEET 4

1. Which of the following in sklearn library is used for hyper parameter tuning?
   A) GridSearchCV() B) RandomizedCV()
   C) K-fold Cross Validation D) None of the above

   A) GridSearchCV()

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

   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

   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?
   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

   C) 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

   A) It's an ensemble of weak learners.

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.

   B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.

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.

D) Both bias and variance decrease.

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?
A)model is underfitting
B) model is overfitting
C) model is performing good
D) None of the above
9.
A)model is underfitting

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

10. 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.

11. What are the advantages of Random Forests over Decision Tree?
1. Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problem in decision trees and also reduces the variance and therefore improves the accuracy.

2. Random Forest can be used to solve both classification as well as regression problems.

3. Random Forest works well with both categorical and continuous variables.

4. Random Forest can automatically handle missing values.

5. No feature scaling required: No feature scaling (standardization and normalization) required in case of Random Forest as it uses rule based approach instead of distance calculation.

6. Handles non-linear parameters efficiently: Non linear parameters don't affect the performance of a Random Forest unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Random Forest may outperform as compared to other curve based algorithms.

7. Random Forest can automatically handle missing values.

8. Random Forest is usually robust to outliers and can handle them automatically.

9. Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.

10. Random Forest is comparatively less impacted by noise.

12. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.
It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.
two techniques used for scaling :
Rank Order
Constant Sum

13. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
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?
A pair of evaluations metrics that are commonly used when there is a class imbalance are precision and recall. Precision is defined as the number of true positives divided by the sum of true positives and false positives.

14. What is “f-score" metric? Write its mathematical formula.
The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall. The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

F1 =2 * (pecision * recall) / (precision + recall)

15. What is the difference between fit(), transform() and fit_transform()?
Fit means to fit the model to the data being provided. This is where the model "learns" from the data.

Transform means to transform the data (produce model outputs) according to the fitted model.

fit_transform means to do both - Fit the model to the data, then transform the data according to the fitted model. Calling fit_transform is a convenience to avoid needing to call fit and transform sequentially on the same input.