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?
   1. GridSearchCV() B) RandomizedCV()



C) K-fold Cross Validation D) None of the above

**Ans: a) GridSearchCV()**

1. In which of the below ensemble techniques trees are trained in parallel?
   1. Random forest B) Adaboost

C) Gradient Boosting D) All of the above

**Ans: A) Random forest**

1. 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?

* 1. The regularization will increase B) The regularization will decrease

C) No effect on regularization D) kernel will be changed to linear

**Ans: A)** **The regularization will increase**

1. 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?

* 1. It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.
  2. It denotes the number of children a node can have.
  3. both A & B
  4. None of the above

**Ans: C) both A & B**

1. Which of the following is true regarding Random Forests?
   1. It's an ensemble of weak learners.
   2. The component trees are trained in series
   3. 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: A) It's an ensemble of weak learners*.***

1. What can be the disadvantage if the learning rate is very high in gradient descent?
   1. Gradient Descent algorithm can diverge from the optimal solution.
   2. Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.
   3. Both of them D)None of them.

**Ans:A) Gradient Descent algorithm can diverge from the optimal solution.**

1. As the model complexity increases, what will happen?
   1. Bias will increase, Variance decrease B) Bias will decrease, Variance increase C) both bias and variance increase D) Both bias and variance decrease.

**Ans: D) Both bias and variance decrease*.***

1. 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?

* 1. model is underfitting B) model is overfitting

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

**Ans: A)** **model is underfitting**

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

1. 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.
2. What are the advantages of Random Forests over Decision Tree?

* Random Forest is based on the bagging algorithm and uses Ensemble Learning techniques. It generates 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 reduces the variance which increases the

accuracy.

* Random Forest can automatically handle missing values. No standardization or normalization is required as Random Forest uses rule-based approach instead of distance calculation.
* It Handles the non-linear parameters efficiently: Nonlinear parameters don't affect the performance of a Random Forest unlike curve-based algorithms. So, if there is high nonlinearity between the independent variables, Random Forest may outperform as compared to other curve-based algorithms.
* It is usually robust to outliers and can handle them automatically. 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.

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

* Scaling is a Data Pre-Processing step which is applied to independent variables or features of

data. It basically treats all the data and keep them within a predefined range.It helps the machine to process the data more efficiently.

The two techniques involved in scaling are

1. Normalization

ii) Standardisation

**Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.**

**Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.**

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

1. 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.

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

* The F-score or 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 \* (precision \* recall) / (precision + recall)

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

* Transform means to transform the data (produce model outputs) according to the model fitted.
* fit\_transform do both the task which mean that it Fit the model to the data, and also transform the data accordingly. Using fit\_transform is better as it helps to avoid call fit and transform sequentially for same input.