 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?
   1. GridSearchCV()
2. In which of the below ensemble techniques trees are trained in parallel?
   1. 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?

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?

* 1. It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.

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

A) It's an ensemble of weak learners.

1. What can be the disadvantage if the learning rate is very high in gradient descent?
2. Gradient Descent algorithm can diverge from the optimal solution.
3. As the model complexity increases, what will happen?

B) Bias will decrease, Variance increase

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?

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.

Gini index = 1 – (0.4^2 + 0.6^2) = 0.48

Entropy = – (0.4\*log0.4 + 0.6\*lo­g0.6) = 0.97

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

Random forest is an ensemble method in which a classifier is constructed by combining several different *Independent*base classifiers. The independence is theoretically enforced by training each base classifier on a training set sampled with replacement from the original training set. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features.

The ensemble classifier then aggregates the individual predictions to combine into a final prediction, based on a majority voting on the individual predictions.

It can be shown that an ensemble of independent classifiers, each with an error rate e, when combined significantly reduces the error rate.

Suppose we have 10 *independent*classifiers, each with error rate of 0.3

ϵ=0.3ϵ=0.3

In this setting, the error rate of the ensemble can be computed as below (we are taking a majority vote on the predictions. An ensemble makes a wrong prediction only when more than half of the base classifiers are wrong)

ϵensemble=∑i=10i=6(10i)ϵi(1−ϵ)10−i≈0.05ϵensemble=∑i=6i=10(10i)ϵi(1−ϵ)10−i≈0.05

It can be seen that with the theoretical guarantees stated above an ensemble model performs significantly well.

However in practice it is not possible to guarantee such classifier independence as they are trained from the same data, but still introduction of randomness helps achieve independence to a certain degree and it has been empirically observed that ensembles perform significantly well over individual base classifiers.

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data

processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Two techniques used for scaling are Min Max Scaler and Standard Scaler.

 WORKSHEET

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

Feature scaling helps in causing Gradient Descent to converge much faster as standardizing all the variables on to the same scale, for example, for a linear regression makes it easy to calculate the slope ( y = mx + c) (where we normalize the M parameter to converge faster).

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?

Unlike standard evaluation metrics like accuracy, that treat all classes as equally important, imbalanced classification problems typically rate classification errors with the minority class as more important than those with the majority class. As such performance metrics may be needed that focus on the minority class, which is made challenging because it is the minority class where we lack observations required to train an effective model.

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

F1 score (also F-score or F-measure) is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.

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

"Fit" computes the mean and std to be used for **later** scaling.

"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.

