Machine Learning WORKSHEET-7
ANSWERS
FlipRobo-INT32

01. D) All of the above
02. A) Random forest
03. B) The regularization will decrease
04. C) both A & B
05. A) It's an ensemble of weak learners
06. C) Both of them
07. B) Bias will decrease, Variance increase
08. B) model is overfitting

09. **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.**

ANS: The Gini index and entropy of the dataset with 40% A and 60% B classes can be calculated as follows:

Gini Index = 1 - (0.4)^2 - (0.6)^2 = 0.48
Entropy = -(0.4log2(0.4) + 0.6log2(0.6)) = 0.97

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

Ans: Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. Decision trees have a low bias and are non-parametric, they suffer from a high variance which makes them less useful for most practical applications. By aggregating multiple decision trees, one can reduce the variance of the model output significantly, thus improving performance. While this could be archived by simple tree bagging, the fact that each tree is build on a bootstrap sample of the same data gives a lower bound on the variance reduction, due to correlation between the individual trees. Random Forest addresses this problem by sub-sampling features, thus de-correlating the trees to a certain extend and therefore allowing for a greater variance reduction / increase in performance.

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

In many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude. Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. Two most common techniques of feature scaling are Normalization and Standardization.

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

ANS: Gradient descent is an optimization algorithm used to minimize the cost function in machine learning algorithms like Logistic Regression, SVM, Neural Networks etc. If features are on different scale, certain weights are updated faster than others in Gradient Descent. However, feature scaling helps in causing Gradient Descent to converge much faster as standardizing all the variables on to the same scale.

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?

ANS: Accuracy Metric is one the simplest and widely used metric to measure the performance of a classification predictive model. The reason for its wide use is because it is easy to calculate, easy to interpret, and is a single number to summarize the model's capability. However, accuracy metric fails to perform on an imbalanced dataset as it gives misleading conclusions. In an imbalanced dataset getting an accuracy score of 90 or 99 are trivial as model might have considered the less numbered observation as error or outliers and could have ignored them in the prediction.

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

ANS: 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'.

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

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

ANS: In machine learning, fit(), transform() and fit_transform() are methods of the scikit-learn library used for preprocessing data:

- fit() method is used to fit the data to the model, it is used to calculate the internal parameters of the model.
- transform() method is used to transform the data according to the internal parameters calculated during the fit() method.
- fit_transform() method is used to f`it the data to the model and then transform it in one step.