ASSIGNMENT – 6

MACHINE LEARNING

1. In which of the following you can say that the model is overfitting?

Answer: C) High R-squared value for train-set and Low R-squared value for test-set.

2. Which among the following is a disadvantage of decision trees?

Ans: B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

Ans: C) Random Forest

4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?

Ans: B) Sensitivity

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?

Ans: B) Model B

6. Which of the following are the regularization technique in Linear Regression??

Ans: A) Ridge

7. Which of the following is not an example of boosting technique?

Ans: C) Random Forest

8. Which of the techniques are used for regularization of Decision Trees?

Ans: A) Pruning

9. Which of the following statements is true regarding the Adaboost technique?

Ans: D) None of the above

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

Ans: The adjusted R-squared is a modified version of the R-squared that takes into account the number of predictors in the model. It penalizes the presence of unnecessary predictors by adjusting the R-squared value based on both the number of predictors and the sample size. The adjusted R-squared penalizes the addition of predictors that do not improve the model's performance, and helps to ensure that the model does not become overfitted. Therefore, it can be used to compare models with different numbers of predictors and can help to identify whether adding additional predictors to the model leads to a significant improvement in the model's performance or not.

11. Differentiate between Ridge and Lasso Regression.

Ans: Principal Component Analysis (PCA) is a widely used technique in data analysis and machine learning that is used to reduce the dimensionality of large datasets. It works by identifying and retaining the most important patterns and relationships within the data by projecting it onto a lower-dimensional space. The goal of PCA is to find a new set of variables (known as principal components) that are uncorrelated and that capture as much of the variance in the original data as possible. This can be useful in cases where the original dataset contains a large number of variables that are highly correlated with each other, making it difficult to identify important patterns or relationships. By reducing the dimensionality of the data, PCA can simplify data analysis and modeling, improve interpretability of results, and reduce the risk of overfitting.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans: Variance Inflation Factor (VIF) measures the degree of multicollinearity in a linear regression model. A high VIF indicates high correlation among the predictor variables and may lead to unstable and unreliable coefficient estimates. A VIF value greater than 5 or 10 is often used as a threshold for high multicollinearity. Addressing multicollinearity issues can involve removing correlated variables, transforming variables, or using regularization techniques like Ridge or Lasso regression.

13. Why do we need to scale the data before feeding it to the train the model?

Ans: Scaling the data is an important preprocessing step in machine learning because it can improve the performance and convergence of many machine learning algorithms. This is because most machine learning algorithms are based on the distance or similarity between data points, and if the features in the dataset are on different scales, this can lead to biased or inconsistent results.

Scaling the data helps to normalize the feature values so that they are all on a similar scale. This can make it easier for the machine learning algorithm to find the optimal solution by ensuring that no one feature dominates the others. Additionally, scaling can also help to speed up the training process and improve the generalization performance of the model by reducing the impact of outliers.

There are several methods for scaling data, including standardization (subtracting the mean and dividing by the standard deviation) and normalization (scaling values to a range of 0 to 1). The specific method chosen will depend on the distribution of the data and the requirements of the machine learning algorithm being used.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Ans: There are several metrics used to check the goodness of fit in linear regression models, including:

1. R-squared (R2) - measures the proportion of variation in the dependent variable that is explained by the independent variables. Higher values of R2 indicate a better fit.
2. Mean Squared Error (MSE) - measures the average squared difference between the actual and predicted values of the dependent variable. Lower values of MSE indicate a better fit.
3. Root Mean Squared Error (RMSE) - the square root of the MSE, which is expressed in the same units as the dependent variable. Like MSE, lower values of RMSE indicate a better fit.
4. Mean Absolute Error (MAE) - measures the average absolute difference between the actual and predicted values of the dependent variable. Lower values of MAE indicate a better fit.
5. Residual plots - graphical representations of the difference between the predicted and actual values of the dependent variable. These plots can be used to identify patterns or trends in the residuals that may indicate a poor fit.