MACHINE LEARNING

ASSIGNMENT - 6

1. In which of the following you can say that the model is overfitting?
   Ans. A) High R-squared value for train-set and High R-squared value for test-set.
2. Which among the following is a disadvantage of decision trees?
   Ans. C) Decision trees are not easy to interpret

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. C) Precision

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 techniques in Linear Regression??

Ans. A) Ridge

D) Lasso

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

Ans. B) Decision Tree

C) Random Forest

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

Ans.D) All of the above

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

Ans. A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

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 R-squared that adjusts for predictors that are not significant in a regression model.

Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11. Differentiate between Ridge and Lasso Regression.

Ans. **Ridge regression** is mostly used to reduce the over fitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.

Lasso regression helps to reduce the over fitting in the model as well as feature selection.

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

Ans. A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

- **A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.**
- **Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.**
- **A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.**

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

Ans. It dramatically improves model accuracy. Normalization gives equal weights/importance to each variable so that no single variable steers model performance in one direction just because they are bigger numbers.

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

Ans. Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE).

The adjusted R-square statistic is generally the best indicator of the fit quality when you add additional coefficients to your model. The adjusted R-square statistic can take on any value less than or equal to 1, with a value closer to 1 indicating a better fit. A RMSE value closer to 0 indicates a better fit.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.