MACHINE LEARNING

ASSIGNMENT - 6

In Q1 to Q5, only one option is correct, Choose the correct option:

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

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?

B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

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?

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?

D) Data Insufficient

In Q6 to Q9, more than one options are correct, Choose all the correct options:

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

A) Ridge

D) Lasso

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

A) Adaboost

D) Xgboost.

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

A) Pruning

C) Restricting the max depth of the tree

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

C) It is example of bagging technique

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

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

11. Differentiate between Ridge and Lasso Regression.

In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The *L2* term is equal to the square of the magnitude of the coefficients. While, in Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function. This term is the absolute sum of the coefficients.

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

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

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

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).

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

|  |  |  |
| --- | --- | --- |
| Actual/predicted | True | False |
| True | 1000 | 50 |
| False | 250 | 1200 |

Sensitivity= (TP)/(TP+FN)= 1000+1250= 0.8

Specificity= (TN)/(FP+TN)=1200/1250= 0.96

Precision= (TP)/(TP+FP)= 1000/1050= 0.95

Accuracy= = (TN+TP)/(TP+FP+TN+FN)=2200/2500=0.88