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?

A) High R-squared value for train-set and High R-squared value for test-set.

B) Low R-squared value for train-set and High R-squared value for test-set.

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

D) None of the above

Ans :- C

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

A) Decision trees are prone to outliers.

B) Decision trees are highly prone to overfitting.

C) Decision trees are not easy to interpret

D) None of the above.

Ans :- B

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

Ans :- A

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?

A) Accuracy B) Sensitivity

C) Precision D) None of the above.

Ans :- C

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?

A) Model A B) Model B

C) both are performing equal D) Data Insufficient

Ans :- B

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 B) R-squared

C) MSE D) Lasso

Ans :- B & C

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

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

Ans :- B & C

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

A) Pruning B) L2 regularization

C) Restricting the max depth of the tree D) All of the above

Ans :- A & C

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

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

C) It is example of bagging technique

D) None of the above

Ans :- A & B

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?

Ans :- It penalizes for adding unnecessary features and allows a comparison of regression models with a different number of predictors. Here k is the number of explanatory variables in the model and n is the number of observations. The value of adjusted R^2 is always less than that of R^2.

11. Differentiate between Ridge and Lasso Regression.

Ans :- Main difference between Ridge and LASSO Regression is that if ridge regression can shrink the coefficient close to 0 so that all predictor variables are retained. Whereas LASSO can shrink the

coefficient to exactly 0 so that LASSO can select and discard the predictor variables that have the right coefficient of 0.

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.

Most statistical software has the ability to compute VIF for a regression model. The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows: A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.

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

Ans :- To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.

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

Ans :- Five metrics give us some hints about the goodness-of-fit of our model. The first two metrics, the Mean Absolute Error and the Root Mean Squared Error (also called Standard Error of the Regression), have the same unit as the original data.

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

Actual/Predicted True False

True 1000 50

False 250 1200

Ans :- Sensitivity = 0.8

Specificity = 0.96

Precision = 0.8

Recall = 0.8

Accuracy = 0.88