Assignment-4

Machine Learning

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:- None of the above

2. Wh ich 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:- Decision trees are highly prone to overfitting

3. Which of the following is an ensemble technique?
   A) SVM B) Logistic Regression
   C) Random Forest D) Decision tree

   Ans:- 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?
   A) Accuracy B) Sensitivity
   C) Precision D) None of the above.

   Ans:- Accuracy

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:- Model A

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:- Ridge and Lasso

7. Which of the following is not an example of boosting technique?
   A) Adaboost B) Decision Tree
   C) Random Forest D) Xgboost.

   Ans:- Decision Tree

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:- pruning

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:- None of the above

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:- 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 and Lasso regression uses two different penalty functions for regularisation. Ridge regression uses L2 on the other hand lasso regression go uses L1 regularisation technique. In ridge regression, the penalty is equal to the sum of the squares of the coefficients and in the Lasso, penalty is considered to be the sum of the absolute values of the coefficients. In lasso regression, it is the shrinkage towards zero using an absolute value (L1 penalty or regularization technique) rather than a sum of squares(L2 penalty or regularization technique)

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

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.

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

Ans:- MSE(Mean square error) RMSE(Root Mean Square Error)or MAE *(*mean absolute error) are better be used to compare performance between different regression models.

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=0.95
Specificity=0.82
Precision=0.8
Recall=0.95
Accuracy=0.88