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?
   - High R-squared value for train-set and High R-squared value for test-set.
   - Low R-squared value for train-set and High R-squared value for test-set.
   - High R-squared value for train-set and Low R-squared value for test-set.
   - None of the above

ANS :- None of the above

2. Which among the following is a disadvantage of decision trees?
   - Decision trees are prone to outliers.
   - Decision trees are highly prone to overfitting.
   - Decision trees are not easy to interpret
   - None of the above.

ANS :- Decision trees are highly prone to overfitting.

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

ANS :- C) Random Forest

4. Suppose you are building a classification model for detection of a fatal disease where detection ofthe 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 :- A)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 :-C) both are performing equal

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 :- A)Ridge & D) Lasso

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

A)Adaboost
B) Decision Tree
C) Random Forest
D) Xgboost.

ANS :-C) Random Forest

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) Pruning & B) L2 regularization

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

- We initialize the probabilities of the distribution as 1/n, where n is the number of data-points
- A tree in the ensemble focuses more on the data points on which the previous tree was notperforming well
- It is example of bagging technique
- 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 themodel?

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

11. Differentiate between Ridge and Lasso Regression.

ANS :- The 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 regressionmodelling?

ANS :- The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity.

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

ANS :- Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem. In the case of neural networks, an independent variable with a spread of values may result in a large loss in training and testing and cause the learning process to be unstable.

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

ANS :-These are the different metrics which are used to check the goodness of fit in linear regression

1.Mean Absolute Error(MAE)

2.Root Mean Square Error(RMSE)

3.Coefficient of determination or R2

4.Adjusted R2

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

Actual/Predicted	True	False
True	1000	50