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: 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? 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) 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: A) SVM

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.

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: B) Model 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: 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: B) Decision Tree, 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

D) All of the above

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

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 the R-squared metric that takes into account the number of predictor variables included in a linear regression model. It adjusts the R-squared value to account for the fact that adding additional predictors to a model will generally increase the R-squared value, even if the additional predictors do not actually improve the model's ability to explain the variance in the dependent variable.
11. Differentiate between Ridge and Lasso Regression.
    Ridge regression and Lasso regression are two popular techniques for regularizing linear regression models to prevent overfitting.

    The main difference between Ridge and Lasso regression lies in the type of penalty term that is added to the regression objective function to constrain the values of the regression coefficients.
12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
    VIF stands for Variance Inflation Factor, which is a measure of the degree of multicollinearity between predictor variables in a regression model. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, which can make it difficult for the regression model to estimate the contribution of each variable to the outcome variable independently.
13. Why do we need to scale the data before feeding it to the train the model?
    machine learning model. The main reasons for scaling the data are:

    - To improve the performance of the model:
    - To speed up the training process:
    - To prevent numerical instability:

MACHINE LEARNING

14. What are the different metrics which are used to check the goodness of fit in linear regression?
    R-Square, RMSE, MSE
15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Accuracy = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.88
Precision = 1000 / (1000 + 250) = 0.80
Recall = 1000 / (1000 + 50) = 0.95
F1-score = 2 * (precision * recall) / (precision + recall) = 0.87