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: B)
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:C)

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

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

MACHINE LEARNING

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 ans 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: 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.
    - Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1.
    - In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values 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?

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

As a rule of thumb, a VIF of three or below is not a cause for concern. As VIF increases, the less reliable your regression results are going to be.

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.

MACHINE LEARNING

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

    Ans: 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

Ans: **Sensitivity = TP/ TP+FN**
1000/1000+250 = 1000/1250 = 0.8

Specificity= TN/ FP+TN
1200/50+1200 = 1200/1250 = 0.96

Precision = TP/TP+FP
1000/1000+50 = 1000/1050 = 0.95

Recall = TP/TP+FN
1000/1000+250 = 1000/1250 =0.8

Accuracy = TP+TN/ TP+FP+TN+FN
1000+1200/ 1000+50+1200+250 = 2200/2500 = 0.88