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
   Answer : 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.
   Answer :- B)

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

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
   Answer :- 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
   Answer : A) and D)

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

Answer: B)

MACHINE LEARNING

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
   Answer :- A)

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?

    Answer :- The adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the R-squared value. The adjusted R-squared takes into account the number of predictors in the model and increases only if the new predictor improves the model more than what would be expected by chance.

11. Differentiate between Ridge and Lasso Regression.
    Answer :- Ridge and Lasso regression are both methods of regularization in linear regression. Ridge regression adds a penalty term to the least squares objective function that is equal to the square of the magnitude of the coefficients. This has the effect of shrinking the coefficients towards zero, but it does not set any coefficients exactly to zero. Lasso regression, on the other hand, adds a penalty term to the least squares objective function that is equal to the absolute value of the magnitude of the coefficients. This can set some coefficients exactly to zero, which is useful for feature selection.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
    Answer :- VIF (Variance Inflation Factor) is a measure of how much the variance of the estimated regression coefficients is increased because of collinearity. A VIF of 1 indicates no collinearity, whereas a VIF greater than 1 indicates increasing collinearity. A suitable value for a VIF for a feature to be included in a regression model is generally considered to be less than 5.

13. Why do we need to scale the data before feeding it to the train the model?
    Answer :- Scaling the data before feeding it to the model is important because many machine learning algorithms use some form of distance to inform them, for example, k-nearest neighbors and Support Vector Machines. If one of the features has a much larger scale than the others, it can dominate the distance measure, leading to poor performance. Scaling the data can help to avoid this issue.

MACHINE LEARNING

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

Answer :- 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

Answer :- sensitivity or recall or True positive rate - 0.952
specifity :- 0.96
precision :-0.8
accuracy :- 0.88