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 :- (C)

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 :- (B)

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,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)

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 :- (D)

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 :- (B)

Q10 to Q15 are subjective answer type questions, Answer them briefly.

9. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

   Answer:- Adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the R-squared value for each additional predictor added to the model. The adjusted R-squared value accounts for the number of predictors in the model and the sample size, and increases only if the addition of a predictor significantly improves the model's ability to predict the response. As a result, models with many unnecessary predictors will have a lower adjusted R-squared value compared to models with only important predictors. This helps in identifying the most important predictors for the response.

10. Differentiate between Ridge and Lasso Regression.

    Answer:- Ridge and Lasso Regression are both regularization techniques used to prevent over fitting in linear regression models. The key differences between them are:

    Penalty Term: Ridge Regression uses L2 regularization, which adds a penalty term equal to the square of the magnitude of the coefficients. Lasso Regression uses L1 regularization, which adds a penalty term equal to the absolute value of the coefficients.

    Feature Selection: Lasso Regression has the ability to perform feature selection by shrinking the coefficients of unimportant predictors to zero, effectively removing them from the model. Ridge Regression does not perform feature selection and all predictors remain in the model.

    Bias-Variance Trade-Off: Ridge Regression tends to have lower bias and higher variance compared to Lasso Regression, which has higher bias and lower variance.

    Model Interpretation: Ridge Regression produces a continuous output, making it easier to interpret the model's coefficients. Lasso Regression, on the other hand, produces sparse models that can be difficult to interpret.

11. 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 statistical measure used to assess the presence of multicollinearity in a linear regression model. It measures the amount of increase in the variance of a predictor's coefficient due to the presence of other correlated predictors in the model.A VIF value of 1 indicates that there is no multicollinearity, while a value greater than 1 indicates that the presence of other predictors is causing an inflation of the variance of the coefficient of the predictor being evaluated. A commonly used threshold for a suitable value of VIF is 10, meaning that if a predictor has a VIF greater than 10, it may indicate multicollinearity and the predictor should be re-evaluated for inclusion in the model. However, the threshold can vary based on the specific requirements and nature of the data.

MACHINE LEARNING

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

Answer:- Scaling the data before training a model is important for several reasons:

Improves model performance: Some machine learning algorithms, such as k-nearest neighbors and support vector machines, are sensitive to the scale of the features and can perform better when the features are scaled to the same range.

Improves optimization: Some optimization algorithms, such as gradient descent, converge faster when the features are scaled to similar ranges.

Prevent biasing: Features with larger scales can dominate the loss function and affect the model's performance, biasing the model towards those features. Scaling the data helps to prevent this bias and ensures that all features contribute equally to the model.

Increases interpretability: Scaling the data can also make it easier to interpret the coefficients of the model, as the scale of the coefficients reflects the importance of each feature in the model.

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

There are several metrics used to evaluate the goodness of fit in linear regression, including:

R-Squared
Mean Absolute Error (MAE)
Mean Squared Error (MSE)
Root Mean Squared Error (RMSE)
Adjusted R-Squared

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

MACHINE LEARNING