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 - High R-squared value for train-set and Low 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.

Answer - 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

Answer - Decision tree

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 – 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

Answer – Model B

FLIP ROBO

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 – Lasso

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

Answer – 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

Answer – Pruning

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 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?

Answer - Investors can compare the performance of a mutual fund to a benchmark using R-squared and modified R-squared. They may also be used by investors to determine how well their portfolio performed in comparison to a certain benchmark. R-squared is a measure of correlation in the investment industry that is stated as a percentage ranging from 0 to 100, with 100 denoting perfect connection and 0 denoting no correlation at all. The number does not reflect the performance of a specific class of securities. It just assesses how closely the returns match those of the benchmark being assessed. It is also retroactive and unreliable as a forecast of outcomes.

11. Differentiate between Ridge and Lasso Regression.
Answer – Below is the difference between ridge and lasso regression.
Lasso regression: Least Absolute Shrinkage and Selection Operator is referred to as Lasso. This approach composes data points towards a central point, such as the mean. Another name for Lasso is L1 regularization. When the model is computationally challenging or overfit, it is used. Lasso employs || to penalize the high coefficients in contrast to Ridge Regression.
The shrinkage factor is calculated by multiplying the regression coefficients by lambda (The last factor in the above equation). Without further ado, let's have a look at a Ridge and Lasso Model construction example! Let's think about the dataset from an advertising agency! They tracked their sales against an advertisement they ran on various media, including TV, radio, and newspapers!
Ridge Regression: Analysis of multi-linear regression (multicollinear), sometimes referred to as L2 regularization, is done using the ridge regression approach. When anticipated values exceed observed values, it is applied. The equation's tuning parameter lambda () is chosen using the cross-validation approach, which shrinks the squares (2) to make the fit smaller.
Regression coefficients' squared sum times lambda is the shrinking factor (The last element in the above equation).

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Answer - One indicator of the degree of multicollinearity in regression analysis is the variance inflation factor (VIF). In a multivariate regression model, multicollinearity occurs when there is a correlation between several independent variables. The regression findings may suffer as a result. The variance inflation factor may therefore be used to calculate the degree to which multicollinearity has inflated the variance of a regression coefficient. Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of 5 or even 2.5.

MACHINE LEARNING

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

Answer - 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.

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

Answer: Adjusted R-squared, The F-test, and RMSE are used to check the goodness of fit in linear regression. Regression models that fit data well provide projected values that closely match the values of the actual data. If there were no good predictor variables, the mean model, which utilizes the mean for every projected value, would typically be employed. Thus, the suggested regression model's fit ought to be superior to the mean model's fit.

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 = TP / TP + FN
Or Recall = 1000/1000+1200
= 0.45

Specificity = TN / TN + FP
= 50/50+1200
= 0.04

Precision = TP / TP + FP
= 1000/1000+250
=0.8

Accuracy = TP + TN / TP + TN + FP + FN
= 1000+50/1000+50+250+1200
= 0.42

FLIP ROBO