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: 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.

Answer: 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

Answer: C) Random Forest

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: C) Precision

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) 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

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) & C)

MACHINE LEARNING

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: A) & C)

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) & 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 themodel?

Answer:

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. It is used to explain the degree to which input variables (predictor variables) the variation of output variables (predicted variables).

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

11. Differentiate between Ridge and Lasso Regression.

Answer:

The difference between ridge and lasso regression is that **Lasso Regression** tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

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

Answer:

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity. A high VIF means that the independent variable associated with it is strongly collinear with the other variables in the model. If the VIF value is >1, it is said that one can proceed with the regression procedure.

MACHINE LEARNING

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

Answer:

Scaling is the process of placing values in the same range or scale such that no variable is controlled by the other. It is a technique used to standardize the independent features present in data in a fixed range. It is used during data pre-processing to deal with highly varying magnitudes, values, or units.

If feature scaling is not performed, a machine learning algorithm would prefer to weight greater values as higher and consider smaller values as lower, regardless of the unit of measurement. 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:

For check the goodness of fit in linear model we are use three different metrics, which is:

a. Root Mean Squared Error (RMSE)
b. Mean Squared Error (MSE)
c. Mean Absolute Error (MAE)

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 (true positives / all actual positives) = TP / (TP + FN)

= 1000/ (1000+250)

= 1000/ 1250

= ***0.8 or 80% Sensitivity***

Specificity (true negatives / all actual negatives) = TN / (TN + FP)

= 1200/ (1200+50)

= 1200/1250

= ***0.96 or 96% Specificity***

Precision (true positives / predicted positives) = TP / (TP + FP)

= 1000/ (1000+50)

= 1000/ 1050

= ***0.9523 or 95.23% Precision***

MACHINE LEARNING

Recall (true positives / all actual positives) = TP / (TP + FN)

= 1000/ (1000+250)

= 1000/ 1250

= ***0.8 or 80% Recall***

Accuracy (all correct / all) = (TP + TN) / (TP + TN + FP + FN)

= (1000+1200) /(1000+1200+50+250)

= 2200 / 2500

= ***0.88 or 88% Accuracy***