MACHINE LEARNING

I. 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) 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: 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 the model for detection of a fatal disease where detection ofthe 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

II. 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) Ridge D) Lasso

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

MACHINE LEARNING

Answer: C) Random Forest B) Decision Tree

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) L2 regularization C) Restricting the max depth of the tree

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) 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 notperforming well
C) It is an example of a bagging technique

III. 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 is a modified version of the R-squared that adjusts for predictors that are not significant in a regression model. 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.

R-squared increases every time you add an independent variable to the model. The R-squared *never* decreases, not even when it's just a chance correlation between variables. A regression model that contains more independent variables than another model can look like it provides a better fit merely because it contains more variables.

When a model contains an excessive number of independent variables and polynomial terms, it becomes overly customized to fit the peculiarities and random noise in your sample rather than reflecting the entire population. Statisticians call this overfitting the model, and it produces deceptively high R-squared values and a decreased capability for precise predictions.

MACHINE LEARNING

11. Differentiate between Ridge and Lasso Regression.

Answer:

Ridge Regression: In Ridge regression, we add a penalty term equal to the coefficient's square. The *L2* term is equal to the square of the magnitude of the coefficients. We also add a coefficient to control that penalty term. In this case, if is zero then the equation is the basic OLS else if then it will add a constraint to the coefficient. As we increase the value of this constraint causes the value of the coefficient to tend towards zero. This leads to both low variance (as some coefficient leads to negligible effect on prediction) and low bias (minimization of coefficient reduces the prediction dependency on a particular variable).

Lasso Regression:

Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds a penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from *0* this term penalizes, and causes the model, to decrease the value of coefficients in order to reduce loss. The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero compared to Ridge, which never sets the coefficient value to absolute zero.

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

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. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

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. Having featured on a similar scale can help the gradient descent converge more quickly towards the minima.

MACHINE LEARNING

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

Answer:

- R Square/ Adjusted R Square
- Mean Square Error/ Root Mean Square Error
- Mean Absolute Error

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 = 1000/1000+250 = 1000/1250 = 0.8(80%)

Specificity:

TN/TN+FP = 1200/1200+50 = 1200/1250 = 0.96(96%)

Precision:

TP/TP+FP = 1000/1000+50 = 1000/1050 = 0.9523(95%)

Recall:

TP/TP +FN = 1000/1000+250 = 1000/1250 = 0.8(80%)

Accuracy:

TP+TN/TP+TN+FP+FN = 1000+1200/1000+1200+50+250 = 2200/2500 = 0.88(88%)