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

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.

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

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

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

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

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

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

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 takes into account the number of predictors in the model and adjusts the R-squared value accordingly. The adjusted R-squared decreases as the number of predictors increases, thus penalizing the presence of unnecessary predictors in the model. The idea behind this is that adding unnecessary predictors to the model will increase the R-squared value, but this increase will be offset by the increase in the number of predictors, resulting in a decreased adjusted R-squared value. So, the

MACHINE LEARNING

adjusted R-squared provides a better measure of model fit than the R-squared when comparing models with different numbers of predictors.

11. Differentiate between Ridge and Lasso Regression.

Answer: Ridge and Lasso Regression are two popular methods for regularizing linear regression models to prevent overfitting.
Ridge Regression adds a penalty term to the linear regression loss function, equal to the square of the magnitude of the coefficients (L2 regularization). This discourages the coefficients from becoming too large, and encourages them to be small and well-behaved.
Lasso Regression also adds a penalty term to the linear regression loss function, but it is equal to the absolute value of the magnitude of the coefficients (L1 regularization). This encourages the coefficients to be exactly equal to zero, effectively performing feature selection by eliminating the least important features.
In summary, Ridge Regression is better suited for situations where all the predictors are needed, whereas Lasso Regression is better suited for situations where a subset of predictors is sufficient.

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

Answer: VIF stands for Variance Inflation Factor, which is a measure of the amount of multicollinearity in a regression model. VIF is calculated for each predictor in the model, and measures how much the variance of the coefficients is increased due to the presence of other predictors.
A VIF of 1 indicates that there is no multicollinearity, and the predictor is not contributing to the multicollinearity in the model. A VIF value greater than 1, but less than 5, is usually considered acceptable for including a feature in a regression model. A VIF value greater than 5, or even 10 in some cases, indicates high multicollinearity and suggests that the feature should be removed from the model.
In conclusion, the suitable value of a VIF for a feature to be included in a regression model is between 1 and 5, depending on the threshold chosen by the analyst. A VIF value greater than 5 indicates high multicollinearity and suggests that the feature should be removed from the model.

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

Answer: Scaling is necessary in machine learning to ensure that all features are on the same scale and to prevent some features from dominating the model due to their large magnitude. Scaling also helps to avoid numerical instability in certain algorithms and can improve the convergence speed.

MACHINE LEARNING

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

Answer: The different metrics used to check the goodness of fit in linear regression are:

1. Mean Squared Error (MSE)
2. Mean Absolute Error (MAE)
3. R-squared
4. Adjusted R-squared
5. Root Mean Squared Error (RMSE)
6. Mean Absolute Percentage Error (MAPE)
7. Normalized Mean Squared Error (NMSE)
8. Coefficient of Determination (R^2).

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 Positive Rate or Recall) = 1000 / (1000 + 250) = 0.8
Specificity (True Negative Rate) = 1200 / (1200 + 50) = 0.96
Precision = 1000 / (1000 + 50) = 0.95
Recall (True Positive Rate or Sensitivity) = 1000 / (1000 + 250) = 0.8
Accuracy = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.88