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?

Ans: The adjusted R-squared penalizes the presence of unnecessary predictors in the model by adjusting the R-squared value for the number of predictors. When the number of predictors increases, the R-squared value also increases even if the predictors are not useful. The adjusted

MACHINE LEARNING

R-squared corrects for this by subtracting a penalty term for each predictor from the R-squared value.

11. Differentiate between Ridge and Lasso Regression.

Ans: Ridge and Lasso are both regularization techniques used in linear regression to prevent overfitting. The main difference between the two is in how they add the penalty term to the cost function. Ridge regression adds a penalty term to the cost function that is the square of the magnitude of the coefficients, while Lasso regression adds a penalty term that is the absolute value of the magnitude of the coefficients.

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

Ans: VIF (Variance Inflation Factor) is a measure of how much the variance of the estimated regression coefficients are increased because of multicollinearity (correlation) among the independent variables. A VIF of 1 indicates no multicollinearity, while a VIF greater than 1 indicates that there is multicollinearity. A suitable value of VIF for a feature to be included in a regression modeling is less than 5.

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

Ans: Scaling the data before feeding it to the model is important because it ensures that all features are on the same scale. This is necessary because the algorithm uses the scale of the feature to assign weight to it, so if the scales are different, the algorithm will assign more weight to the feature with a larger scale. This can cause the algorithm to be skewed and perform poorly. Additionally, many machine learning algorithms, such as those used in neural networks, require input data to be in a specific range, such as -1 to 1, which is achieved by scaling the data.

MACHINE LEARNING

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

Ans: Different metrics used to check the goodness of fit in linear regression are: • R-squared: a measure of how well the model fits the data. It ranges between 0 and 1, where 1 indicates a perfect fit. • Mean Squared Error (MSE): a measure of the average of the squared differences between the predicted and actual values. • Root Mean Squared Error (RMSE): the square root of the MSE, which gives the same unit of measurement as the data. • Mean Absolute Error (MAE): a measure of the average of the absolute differences between the predicted and actual values.

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans: Sensitivity (True Positive Rate) = 1000 / (1000 + 50) = 0.95

- Specificity (True Negative Rate) = 1200 / (1200 + 250) = 0.82
- Precision = 1000 / (1000 + 250) = 0.8
- Recall (Sensitivity) = 1000 / (1000 + 50) = 0.95
- Accuracy = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.8