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.

MACHINE LEARNING

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. However, the R-squared value can be misleading if the model contains too many unnecessary predictors, which can result in overfitting.

Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictor in the model. It penalizes the presence of unnecessary predictors by reducing the R-squared value if a predictor does not significantly improve the model's fit. This means that the adjusted R-squared value will always be lower than the R-squared value if the model contains more than one predictor.

The formula for adjusted R-squared is:

Adjusted R-squared = 1 - [(1 - R-squared)*(n - 1)/(n - k - 1)]

Where n is the sample size and k is the number of predictors in the model.

As the number of predictors (k) increases, the denominator of the above formula also increases, which leads to a smaller value of adjusted R-squared. This is because the model is penalized for including more predictors that do not significantly improve the model's fit. Therefore, adjusted R-squared is a more appropriate measure than R-squared when comparing models with different numbers of predictors, and it helps to avoid overfitting by penalizing the inclusion of unnecessary predictors in the model.

11. Differentiate between Ridge and Lasso Regression.

Ridge Regression:

Ridge regression adds a penalty term to the cost function of the linear regression model. The penalty term is proportional to the square of the magnitude of the coefficients. The Ridge regression method is also called L2 regularization.

Lasso Regression:

Lasso regression, also called L1 regularization, adds a penalty term to the cost function of the linear regression model, which is proportional to the absolute value of the coefficients. The penalty term is also known as the L1 penalty.

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

VIF stands for Variance Inflation Factor, which is a measure of multicollinearity among the predictor variables in a regression model. Multicollinearity occurs when the predictor variables in a regression model are highly correlated with each other. This makes it difficult for the model to identify the independent effect of each predictor variable on the target variable.

VIF measures the degree to which the variance of the estimated regression coefficient is increased due to multicollinearity. A high VIF value indicates that the corresponding predictor variable is highly correlated with other predictor variables in the model, and its contribution to the model is redundant. A VIF value of 1 indicates no multicollinearity, while a value above 1 indicates the presence of multicollinearity.

MACHINE LEARNING

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

Improving Model Performance:

Many machine learning algorithms use some form of distance calculation, such as Euclidean distance or cosine similarity, to determine the similarity between data points. If the variables are on different scales, variables with larger values will dominate the distance calculations, and variables with smaller values will have a minimal impact. Scaling the data to a common scale ensures that all variables contribute equally to the distance calculations, which can improve the performance of the model.

Gradient Descent:

Many machine learning algorithms use gradient descent to optimize the parameters of the model. Gradient descent algorithms converge faster when the variables are on a similar scale, and scaling the data can reduce the number of iterations required to reach convergence.

Regularization:

Regularization techniques, such as Ridge and Lasso regression, penalize large coefficients in the model. If the variables are on different scales, the regularization penalty will be more significant for variables with larger values. Scaling the data ensures that the regularization penalty is applied equally to all variables.

MACHINE LEARNING

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

R-squared (R^2): It measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. A higher R^2 value indicates a better fit.

Mean Squared Error (MSE): It measures the average squared difference between the predicted and actual values. A lower MSE value indicates a better fit.

Root Mean Squared Error (RMSE): It measures the square root of the MSE. It is in the same units as the dependent variable, and a lower value indicates a better fit.

Mean Absolute Error (MAE): It measures the absolute difference between the predicted and actual values. It is in the same units as the dependent variable, and a lower value indicates a better fit.

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

From the given confusion matrix:

True Positive (TP) = 1000
False Positive (FP) = 50
False Negative (FN) = 250
True Negative (TN) = 1200
Using these values, we can calculate the different metrics:

Sensitivity = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8 or 80%
Specificity = TN / (TN + FP) = 1200 / (1200 + 50) = 0.96 or 96%
Precision = TP / (TP + FP) = 1000 / (1000 + 50) = 0.952 or 95.2%
Recall = Sensitivity
Accuracy = (TP + TN) / (TP + FP + FN + TN) = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.892 or 89.2%