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?

    The adjusted R-squared is a modified version of the R-squared value that takes into account the number of predictors included in a linear regression model. It penalizes the presence of unnecessary predictors in the model by adjusting the R-squared

MACHINE LEARNING

value to account for the number of predictors used. The adjusted R-squared is calculated as: Adjusted R-squared = 1 - [(1 - R-

squared) * (n - 1) / (n - k - 1)] where n is the number of observations in the data set, k is the number of predictors in the model, and R-squared is the coefficient of determination, which measures the proportion of variance in the dependent variable that is explained by the independent variables. The adjusted R-squared penalizes the inclusion of unnecessary predictors in the model by reducing the value of R-squared when additional predictors are added that do not improve the overall fit of the model. This is because adding unnecessary predictors to the model increases the denominator of the equation, (n - k - 1), which reduces the adjusted R-squared value. By penalizing the presence of unnecessary predictors in the model, the adjusted R-squared provides a more accurate measure of the goodness of fit of the model, and helps to prevent overfitting, which can occur when too many predictors are included in the model.

11. Differentiate between Ridge and Lasso Regression.

Ridge regression and Lasso regression are two common techniques used for linear regression analysis with high-dimensional data or when multicollinearity is present. They both aim to address the issue of overfitting in a linear regression model by introducing a penalty term in the cost function that shrinks the coefficients of the predictors towards zero. However, there are some differences between the two techniques: Type of penalty: The main difference between Ridge and Lasso regression is the type of penalty used. Ridge regression uses L2 regularization, which adds the sum of the squared values of the coefficients to the cost function, while Lasso regression uses L1 regularization, which adds the sum of the absolute values of the coefficients to the cost function. Effect on coefficients: Ridge regression shrinks the coefficients towards zero, but does not set them exactly to zero, while Lasso regression can set some of the coefficients to exactly zero, effectively performing feature selection by removing some of the predictors from the model. Complexity: Ridge regression is simpler to implement and computationally less intensive than Lasso regression. Lasso regression requires more computational resources because it involves an optimization problem that requires the use of nonlinear optimization algorithms. Bias-variance trade-off: Ridge regression reduces the variance of the model, but may increase the bias, while Lasso regression reduces both the variance and the bias of the model. In summary, while both Ridge and Lasso regression are useful techniques for addressing the issue of overfitting in linear regression models, they have different strengths and weaknesses and should be chosen based on the specific characteristics of the data and the goals of the analysis. Ridge regression may be preferred when all predictors are thought to be relevant, while Lasso regression may be preferred when there are many predictors and some are expected to be irrelevant.

MACHINE LEARNING

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 the extent of multicollinearity in a regression model. Multicollinearity occurs when two or more predictors in a regression model are highly correlated, which can lead to unstable and unreliable estimates of the coefficients of the predictors. The VIF is calculated for each predictor in the model by regressing it against all the other predictors in the model, and then calculating the ratio of the variance of the coefficient estimate to the variance of the coefficient estimate if the predictor was uncorrelated with the other predictors. A VIF of 1 indicates no correlation between the predictor and the other predictors, while a VIF greater than 1 indicates some degree of correlation. A VIF of 5 or greater is often considered an indication of significant multicollinearity, although the threshold may vary depending on the context and the goals of the analysis. In general, it is recommended to remove or combine predictors with high VIF values to improve the stability and interpretability of the regression model. A suitable value of VIF for a feature to be included in a regression model depends on the specific characteristics of the data and the goals of the analysis. However, a common rule of thumb is to use a cutoff value of 2.5 or 3, which indicates moderate correlation between the predictor and the other predictors in the model. It is important to note that the interpretation of VIF should be done in conjunction with other diagnostic measures of multicollinearity, such as the correlation matrix and the eigenvalues of the predictor matrix. Additionally, it is recommended to use a combination of statistical and practical criteria to decide which predictors to include in the model, rather than relying solely on the VIF values.

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

Scaling the data before feeding it to train the model is often an important preprocessing step in many machine learning algorithms. There are several reasons why scaling is important:

i. Different scales of variables: When the variables in the dataset are measured on different scales, it can cause issues for some machine learning algorithms, especially those that are distance-based, such as k-nearest neighbors and support vector machines. Scaling the variables to a common scale ensures that each variable contributes equally to the analysis and avoids the issue of some variables having more weight than others.

ii. Convergence of optimization algorithms: Many machine learning algorithms use optimization algorithms to minimize the objective function, such as the cost function in linear regression.

MACHINE LEARNING

Scaling the variables ensures that the optimization algorithm converges faster and more reliably to the optimal solution. Without scaling, the optimization algorithm may take longer to converge, or it may get stuck in local optima, which can lead to poor performance of the model.

iii. Interpretation of coefficients: In linear models, the coefficients of the predictors represent the change in the response variable for a one-unit increase in the predictor, holding all other predictors constant. If the predictors are not on the same scale, it becomes difficult to interpret the coefficients and compare the relative importance of the predictors.

iv. Regularization: Scaling is important when using regularization techniques, such as L1 or L2 regularization, which penalize the coefficients of the predictors based on their magnitude. Scaling ensures that the regularization penalty is applied fairly to each predictor, regardless of its scale.

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

There are several metrics that are used to check the goodness of fit in linear regression, including:

i. R-squared (R2): R-squared is a measure of the proportion of the variation in the response variable that is explained by the linear regression model. It ranges from 0 to 1, with a value of 1 indicating a perfect fit of the model to the data.

ii. Adjusted R-squared: Adjusted R-squared is a modified version of R-squared that takes into account the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not improve the fit of the model.

iii. Mean squared error (MSE): MSE is a measure of the average squared difference between the predicted values and the actual values of the response variable. It is calculated by taking the sum of the squared residuals and dividing by the number of observations.

iv. Root mean squared error (RMSE): RMSE is the square root of the MSE and is often used to measure the standard deviation of the residuals.

v. Mean absolute error (MAE): MAE is a measure of the average absolute difference between the predicted values and the actual values of the response variable.

vi. Residual standard error (RSE): RSE is a measure of the standard deviation of the residuals and is calculated by taking the square root of the residual sum of squares divided by the degrees of freedom.

MACHINE LEARNING

vii. F-test: The F-test is a statistical test that compares the fit of the full model with the fit of a reduced model that includes only the intercept. It tests the hypothesis that all the coefficients of the predictors in the full model are equal to zero, indicating that the predictors do not contribute significantly to the model.

These metrics are used to evaluate the performance of the linear regression model and to compare different models to determine the best fit for the data.

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, we can calculate the following performance metrics:

Sensitivity (also called recall or true positive rate): the proportion of actual positives that are correctly identified by the model.
Sensitivity = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8

Specificity: the proportion of actual negatives that are correctly identified by the model.
Specificity = True Negatives / (True Negatives + False Positives) = 1200 / (1200 + 50) = 0.96

Precision: the proportion of predicted positives that are correctly identified by the model.
Precision = True Positives / (True Positives + False Positives) = 1000 / (1000 + 50) = 0.952

Recall (also called sensitivity): the proportion of actual positives that are correctly identified by the model.
Recall = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8

Accuracy: the proportion of correct predictions out of all predictions made by the model.
Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + False Negatives + True Negatives) = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.888

Therefore, the sensitivity is 0.8, specificity is 0.96, precision is 0.952, recall is 0.8, and accuracy is 0.888.