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 themodel?**

    Answer - The adjusted R-squared is a modified version of the R-squared value that adjusts for the number of predictors in the model. It is calculated using the following formula:

MACHINE LEARNING

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

where:

R-squared is the original R-squared value of the model.

n is the number of samples in the dataset.

k is the number of predictors in the model.

The adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the value of R-squared as the number of predictors increases. It does this by subtracting a penalty term from the original R-squared value based on the number of predictors in the model.

The intuition behind this is that adding unnecessary predictors to the model may increase the R-squared value, but it will not necessarily improve the model's predictive performance. Instead, it may lead to overfitting and poor generalization performance.

By penalizing the R-squared value for the presence of unnecessary predictors, the adjusted R-squared value provides a more accurate measure of the model's predictive performance and helps to prevent overfitting. It also allows us to compare models with different numbers of predictors and select the one with the best predictive performance while avoiding overfitting.

11. Differentiate between Ridge and Lasso Regression.

Answer – Ridge and Lasso Regression are two popular regularization techniques used in linear regression to prevent overfitting. The key differences between the two techniques are as follows:

1. Penalty function: Ridge regression adds a penalty term to the sum of squares of the coefficients, whereas Lasso regression adds a penalty term to the absolute value of the coefficients.
2. Shrinkage: Ridge regression shrinks the coefficient estimates towards zero, but they are never exactly zero, whereas Lasso regression can shrink some coefficients to zero, effectively performing variable selection and yielding a sparse model.
3. Solution uniqueness: Ridge regression has a unique solution for any given dataset, whereas Lasso regression may have multiple solutions due to the absolute value penalty term.
4. Interpretability: The coefficients obtained from Ridge regression are not easy to interpret, whereas the coefficients obtained from Lasso regression can be interpreted as the importance of the corresponding feature in the model.
5. Handling multicollinearity: Ridge regression handles multicollinearity (high correlation between predictor variables) well by shrinking the coefficients towards zero, whereas Lasso regression may arbitrarily select one of the correlated variables and shrink its coefficient to zero, effectively removing it from the model.

In summary, Ridge regression is a good choice when all predictors are important and multicollinearity is present, while Lasso regression is a good choice when there are many predictors and some of them are not important. Lasso regression can perform variable selection, resulting in a more interpretable and sparse model, but may have multiple solutions and may not handle multicollinearity as well as Ridge regression.

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

Answer – VIF (Variance Inflation Factor) is a measure of the degree of multicollinearity (high correlation between predictor variables) in a linear regression model. It measures

MACHINE LEARNING

how much the variance of the estimated regression coefficient is increased due to multicollinearity in the data.

A VIF value of 1 indicates no multicollinearity, whereas a value greater than 1 indicates the presence of multicollinearity. As a rule of thumb, a VIF value greater than 5 or 10 is considered high and indicates that the corresponding feature is highly correlated with other features in the model. In such cases, it may be necessary to consider removing the feature or combining it with other related features to reduce multicollinearity.

In general, a VIF value less than 5 is considered acceptable for inclusion of a feature in a regression model. However, the specific threshold for a suitable VIF value may vary depending on the domain knowledge, the context and the purpose of the analysis. It is important to check the VIF values of all features in the model to ensure that multicollinearity is not present, as it can lead to unreliable coefficient estimates and reduced predictive power of the model.

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

Answer- Scaling the data is an important preprocessing step in machine learning. Here are some reasons why we need to scale the data before feeding it to the model:

1. **To standardize the range of features**: Different features in a dataset may have different scales and ranges. For example, the values of height and weight are on completely different scales. If we do not scale the data, features with large values will dominate and influence the model more, which can lead to biased predictions. Scaling the data ensures that all features have a similar range and avoids this issue.
2. **To improve model performance**: Some algorithms, such as K-nearest neighbors (KNN) and support vector machines (SVM), are distance-based algorithms that calculate distances between data points. Scaling the data helps these algorithms to work effectively and accurately, as it reduces the effect of differences in feature scales on the distance metric.
3. **To speed up training**: Scaling the data can reduce the time required for model training. Many optimization algorithms, such as gradient descent, converge faster when the features are scaled.
4. **To normalize the data**: Scaling the data can normalize the distribution of features and make it easier to detect outliers, which may have a significant impact on the model performance.

Overall, scaling the data helps to improve the performance and stability of machine learning models and enables us to build better and more accurate models.

MACHINE LEARNING

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

Answer- There are several metrics used to check the goodness of fit in linear regression. Some of the most common ones are:

1. **R-squared (R^2):** This is the most widely used metric to evaluate the goodness of fit in linear regression. It measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. R-squared ranges from 0 to 1, with higher values indicating a better fit.
2. **Mean Squared Error (MSE):** This metric measures the average squared difference between the predicted and actual values of the dependent variable. It gives an idea of how well the model is fitting the data points.
3. **Root Mean Squared Error (RMSE):** This is the square root of the MSE and is used to measure the average difference between the predicted and actual values in the same units as the dependent variable. It is a popular metric because it is easy to interpret and can be directly compared to the scale of the dependent variable.
4. **Mean Absolute Error (MAE):** This metric measures the average absolute difference between the predicted and actual values of the dependent variable. It is less sensitive to outliers than MSE and RMSE.
5. **Adjusted R-squared:** This metric is an adjusted version of R-squared that takes into account the number of independent variables in the model. It penalizes the addition of unnecessary variables to the model and helps to avoid overfitting.

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Answer - Using the values in the confusion matrix, we can calculate the following performance metrics:

Sensitivity (True Positive Rate or Recall): TP / (TP + FN) = 1000 / (1000 + 250) = 0.8
Specificity (True Negative Rate): TN / (TN + FP) = 1200 / (1200 + 50) = 0.96
Precision: TP / (TP + FP) = 1000 / (1000 + 50) = 0.95
Recall (Same as Sensitivity): TP / (TP + FN) = 1000 / (1000 + 250) = 0.8
Accuracy: (TP + TN) / (TP + TN + FP + FN) = (1000 + 1200) / (1000 + 1200 + 250 + 50) = 0.88

Therefore, the sensitivity, specificity, precision, recall and accuracy for the given confusion matrix are 0.8, 0.96, 0.95, 0.8, and 0.88, respectively.