ASSIGNMENT - 6

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

Ans. C

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.

ANS B)

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

ANS. C)

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.

ANS. B

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

ANS. B

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

ANS. A and D

7. Which of the following is not an example of boosting technique?

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

ANS. B and C

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

ANS. A and C

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

ANS . A and B

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 is a modification of the R-squared statistic that takes into account the number of predictors in a regression model. The R-squared is a measure of how well the regression model fits the data, and it ranges from 0 to 1, with higher values indicating a better fit.

When a regression model has more predictors, it is often the case that the R-squared will increase, even if the new predictors do not add any meaningful explanatory power to the model. This is because the R-squared measures the total variation in the response variable that is explained by the model, regardless of whether the variation is due to useful predictors or to noise.

The adjusted R-squared penalizes the presence of unnecessary predictors by adjusting the R-squared for the number of predictors in the model. The adjustment is based on the degrees of freedom, which are the number of observations minus the number of parameters estimated in the model.

The formula for the adjusted R-squared is:

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

where n is the number of observations and p is the number of predictors. The adjusted R-squared will always be less than or equal to the R-squared, and it will decrease as the number of predictors increases, unless the new predictors improve the model fit by a substantial amount.

Therefore, the adjusted R-squared penalizes the presence of unnecessary predictors by reducing the R-squared when additional predictors are added to the model that do not provide significant improvement to the model fit. This helps to prevent overfitting, which occurs when a model is too complex and captures noise in the data rather than the underlying relationships between the predictors and the response variable.

11. Differentiate between Ridge and Lasso Regression.

Ans. Ridge regression and Lasso regression are two types of regularized linear regression techniques used to deal with multicollinearity in the data, where the predictors are highly correlated. The main difference between Ridge and Lasso regression is the way they penalize the coefficients of the predictors.

Ridge Regression:

Ridge regression adds a penalty term to the least squares objective function, which is the sum of the squared differences between the predicted and actual response values. The penalty term is a multiple of the sum of the squares of the coefficients, multiplied by a tuning parameter λ. This penalty term helps to reduce the size of the coefficients and shrinks them towards zero, but it does not set any of them exactly to zero. Ridge regression is particularly useful when all of the predictors are expected to be important, but some of them may have small effects.

Lasso Regression:

Lasso regression also adds a penalty term to the least squares objective function, but the penalty term is the sum of the absolute values of the coefficients, multiplied by a tuning parameter λ. This penalty term not only shrinks the coefficients towards zero but can also set some of them exactly to zero, effectively selecting a subset of the predictors. Lasso regression is particularly useful when there are many predictors, but only a few are expected to be important. Lasso regression can be used for feature selection, as it identifies the most important predictors and eliminates the irrelevant ones.

In summary, Ridge regression tends to shrink the coefficients towards zero without setting them exactly to zero, whereas Lasso regression can set some of the coefficients exactly to zero, effectively eliminating the corresponding predictors. Ridge regression is best suited when all the predictors are important, whereas Lasso regression is preferred when some of the predictors are irrelevant and feature selection is necessary.

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

Ans. VIF stands for Variance Inflation Factor, which is a measure of the amount of multicollinearity in a regression model. Multicollinearity occurs when there is a high correlation between two or more predictors in a regression model, which can lead to unstable and unreliable estimates of the regression coefficients. VIF measures the degree to which the variance of an estimated regression coefficient is increased due to multicollinearity in the data.

The VIF for a predictor is calculated as the ratio of the variance of its estimated coefficient in the full model to the variance of its estimated coefficient in a model that uses only that predictor and the other predictors as a group. 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 common rule of thumb is that a VIF of 10 or greater indicates high multicollinearity and a potential problem in the model. In general, it is recommended to exclude predictors with a VIF greater than 5 or 6, as they can have a significant impact on the estimates of the regression coefficients and the overall model performance. However, the suitable value of VIF for a feature to be included in a regression model ultimately depends on the context and the goals of the analysis.

It is important to note that VIF should not be used as the sole criterion for deciding whether or not to include a predictor in a regression model. Other factors, such as the theoretical importance of the predictor, its statistical significance, and its contribution to the overall model fit, should also be taken into consideration when making this decision. Additionally, VIF is not a reliable measure when the sample size is small, as it can be affected by high variability in the estimates of the regression coefficients.

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

ANS. Scaling the data is an important preprocessing step that is often necessary before training machine learning models. Here are some reasons why we need to scale the data:

A) Different units: If the features of the dataset are measured in different units, such as meters, kilograms, and seconds, then they may have different scales. This can lead to biased estimates of the regression coefficients, as the model will give more weight to features with larger scales. Scaling the data ensures that all features are on the same scale and are equally important in the model.

B) Faster convergence: Many machine learning algorithms use some form of gradient descent optimization, which involves taking small steps in the direction of the steepest descent. Scaling the data can speed up the convergence of the algorithm by reducing the number of iterations required for convergence.

C) Improved model performance: Scaling the data can often lead to improved model performance, as it helps to reduce the impact of outliers and make the model more robust to noise in the data. It can also improve the accuracy and stability of the estimates of the regression coefficients, as it reduces the correlations between the predictors.

D) Regularization: Some regularization techniques, such as L1 regularization (Lasso), require that the features are on the same scale to work effectively. Scaling the data ensures that the regularization is applied equally to all features, which can improve the performance of the model.

There are several ways to scale the data, including standardization, normalization, and min-max scaling, and the choice of scaling method depends on the nature of the data and the requirements of the machine learning algorithm being used.

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

ANS. There are several metrics that are commonly used to check the goodness of fit of a linear regression model:

A) R-squared (R^2): R-squared is a measure of how well the model fits the data. It is the proportion of the total variation in the response variable that is explained by the model. An R-squared value of 1 indicates a perfect fit, while a value of 0 indicates that the model does not explain any of the variation in the response variable.

B) Mean squared error (MSE): MSE measures the average squared difference between the predicted and actual values of the response variable. It is commonly used to evaluate the accuracy of the model's predictions.

C) Root mean squared error (RMSE): RMSE is the square root of the MSE and is also used to evaluate the accuracy of the model's predictions. It is often preferred over MSE because it is in the same units as the response variable.

D) Mean absolute error (MAE): MAE measures the average absolute difference between the predicted and actual values of the response variable. It is also commonly used to evaluate the accuracy of the model's predictions.

E) Residual plots: Residual plots are graphical representations of the differences between the predicted and actual values of the response variable. A good linear regression model will have a pattern-free residual plot, indicating that the residuals are randomly scattered around the zero line.

These metrics can be used together to evaluate the goodness of fit of a linear regression model. R-squared, MSE, RMSE, and MAE are all quantitative measures that provide insight into the accuracy and precision of the model's predictions, while residual plots provide a visual representation of the goodness of fit.

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

Actual/Predicted	True	False
True	1000	50

False	250	120

ANS. Based on the given confusion matrix, the values for the various performance metrics can be calculated as follows:

True Positives (TP) = 1000

False Positives (FP) = 50

False Negatives (FN) = 250

True Negatives (TN) = 1200

1. Sensitivity = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8 or 80%

2. Specificity = TN / (TN + FP) = 1200 / (1200 + 50) = 0.96 or 96%

3. Precision = TP / (TP + FP) = 1000 / (1000 + 50) = 0.95 or 95%

4. Recall = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8 or 80%

5. Accuracy = (TP + TN) / (TP + TN + FP + FN) = (1000 + 1200) / (1000 + 1200 + 50 + 250) = 0.893 or 89.3%

Therefore, the sensitivity is 80%, the specificity is 96%, the precision is 95%, the recall is 80%, and the accuracy is 89.3%.