FLIP ROBO

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?

C) High R-squared value for train-set and Low R-squared value for test-set.

2. Which among the following is a disadvantage of decision trees?
B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?
C) Random Forest

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?
B) Sensitivity

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?
B) Model 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
D) Lasso

7. Which of the following is not an example of boosting technique?
B) Decision Tree

8. Which of the techniques are used for regularization of Decision Trees?
D) All of the above

9. Which of the following statements is true regarding the Adaboost technique?
B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

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 metric that takes into account the number of predictors in the model. It penalizes the presence of unnecessary predictors in the model by adjusting the R-squared value based on the number of predictors in the model.

R-squared is a metric that measures the proportion of the variation in the dependent variable that is explained by the independent variables in the model. However, it does not take into

account the number of predictors in the model. As more predictors are added to the model, the R-squared value may increase, even if the new predictors are not actually adding any significant value to the model.

The adjusted R-squared is calculated as:

Adjusted R-squared = 1 - [(1 - R^2)(n - 1) / (n - k - 1)]

where R^2 is the standard R-squared value, n is the number of observations, and k is the number of predictors in the model.

The adjusted R-squared penalizes the presence of unnecessary predictors in the model by subtracting a penalty term that is proportional to the number of predictors in the model. The penalty term is calculated as [(n - 1) / (n - k - 1)], which increases as the number of predictors in the model increases. This means that as more predictors are added to the model, the adjusted R-squared value will increase at a slower rate than the standard R-squared value, and may even decrease if the new predictors are not adding any significant value to the model.

In summary, the adjusted R-squared penalizes the presence of unnecessary predictors in the model by adjusting the R-squared value based on the number of predictors in the model. It provides a more accurate measure of the model's performance by taking into account the trade-off between the goodness of fit and the complexity of the model.

11. Differentiate between Ridge and Lasso Regression.

Ans - Ridge and Lasso regression are both regularization techniques used in linear regression to prevent overfitting of the model. However, they differ in the way they penalize the coefficients of the regression equation.

Ridge Regression:

In Ridge regression, the sum of the squared values of the coefficients is added to the least squares cost function. This additional term is called the L2 regularization term and is multiplied by a regularization parameter alpha. Ridge regression shrinks the coefficients of less important features towards zero but does not set them to zero. As a result, Ridge regression can be used to reduce the impact of multicollinearity, i.e., the presence of highly correlated predictors in the model.

Lasso Regression:

In Lasso regression, instead of the sum of the squared values of the coefficients, the sum of the absolute values of the coefficients is added to the least squares cost function. This additional term is called the L1 regularization term and is also multiplied by a regularization parameter alpha. Unlike Ridge regression, Lasso regression sets the coefficients of less important features to zero. Therefore, Lasso regression can be used for feature selection, i.e., identifying the most important predictors in the model.

Here are some key differences between Ridge and Lasso regression:

Ridge regression shrinks the coefficients of less important features towards zero, but does not set them to zero, while Lasso regression sets the coefficients of less important features to zero.

Ridge regression can be used to reduce the impact of multicollinearity, while Lasso regression can be used for feature selection.

Ridge regression can be used when all the features are relevant, while Lasso regression can be used when there are many irrelevant features in the model.

The choice between Ridge and Lasso regression depends on the specific problem at hand, the type of data and the goals of the analysis.

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 metric used to detect multicollinearity among the independent variables in a regression model. Multicollinearity is a condition where two or more independent variables are highly correlated with each other, which can lead to inaccurate and unstable regression coefficients.

The VIF of each independent variable is calculated by regressing that variable against all the other independent variables in the model. VIF measures how much the variance of the estimated regression coefficient is increased due to multicollinearity.

The formula for VIF is:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination of the regression of the independent variable on all the other independent variables in the model.

The VIF value of 1 indicates that there is no multicollinearity, while VIF values greater than 1 indicate the presence of multicollinearity. A general rule of thumb is that a VIF value greater than 5 or 10 indicates high multicollinearity, which means that the corresponding independent variable may need to be removed from the model.

Therefore, a suitable VIF value for a feature to be included in a regression modeling depends on the specific problem at hand and the level of tolerance for multicollinearity. However, a VIF value of 5 or less is generally considered acceptable in most cases. It is important to note that the decision to remove a feature based on its VIF value should be taken after considering the importance and relevance of the feature to the problem being solved.

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

Ans Scaling the data is an important step in many machine learning algorithms, especially those that involve distance-based metrics or regularization. Here are some reasons why we need to scale the data before feeding it to the model:

Helps in gradient descent optimization: Many optimization algorithms such as gradient descent converge faster when the features are scaled, as they tend to take longer to converge when the features are on different scales.

Improves numerical stability: Scaling can help prevent numerical instabilities that can arise in certain algorithms when one feature has a much larger magnitude than the others.

Helps with regularization: Regularization techniques such as L1 and L2 regularization depend on the magnitude of the coefficients. Scaling can help ensure that all features are penalized equally.

Helps with distance-based algorithms: Many machine learning algorithms such as K-nearest neighbors and clustering algorithms use distance metrics to measure similarity between samples. If the features are not scaled, the distance metric may be dominated by a feature with a larger magnitude, leading to suboptimal results.

Therefore, scaling the data before feeding it to the model can help improve the performance and stability of the algorithm, and lead to more accurate predictions. There are different scaling techniques such as StandardScaler, MinMaxScaler, MaxAbsScaler, RobustScaler, etc., which can be applied based on the type of data and algorithm used.

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

Ans - In linear regression, there are several metrics that are used to evaluate the goodness of fit of the model. Some of the commonly used metrics are:

R-squared (R^2): This metric measures the proportion of variance in the dependent variable that is explained by the independent variables. R-squared values range from 0 to 1, with higher values indicating a better fit. However, R-squared is not always a reliable measure of model fit, especially when dealing with complex models.

Adjusted R-squared: This metric adjusts for the number of predictors in the model and provides a more accurate measure of model fit than R-squared.

Mean squared error (MSE): This metric measures the average squared difference between the predicted and actual values of the dependent variable. A lower MSE indicates a better fit.

Root mean squared error (RMSE): This metric is the square root of the MSE and provides a measure of the average magnitude of the error in the dependent variable. Lower values of RMSE indicate a better fit.

Mean absolute error (MAE): This metric measures the average absolute difference between the predicted and actual values of the dependent variable. A lower MAE indicates a better fit.

Residual standard error (RSE): This metric measures the standard deviation of the residuals (the difference between the predicted and actual values of the dependent variable). A lower RSE indicates a better fit.

Overall, the choice of metric(s) to use for evaluating the goodness of fit in linear regression depends on the specific goals of the analysis and the nature of the data being modeled. It is often recommended to use a combination of metrics to get a more comprehensive understanding of the model's performance.

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

Ans- In the given confusion matrix:

True positives (TP) = 1000

False positives (FP) = 50

False negatives (FN) = 250

True negatives (TN) = 1200

Using these values, we can calculate the following metrics:

Sensitivity (recall) = 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 (same as sensitivity) = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8 or 80%

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (1000 + 1200) / (1000 + 1200 + 50 + 250) = 0.9 or 90%

Therefore, the sensitivity (recall) is 0.8 or 80%, specificity is 0.96 or 96%, precision is 0.952 or 95.2%, recall is 0.8 or 80%, and accuracy is 0.9 or 90%.