MACHINE LEARNING ASSIGNMENT – 6

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 High R-Squared value for train-set and low R-squared value for the test set

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 Decision tree are highly prone to overfitting.

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression C) Random Forest D) Decision tree

Ans 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?

A) Accuracy B) Sensitivity C) Precision D) None of the above.

Ans

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

Ans Ridge and Lasso

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

A) Adaboost B) Decision Tree C) Random Forest D) Xgboost.

Ans Decision tree and random forest

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 Pruning and restricting the max depth of the tree

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 we initialize the probability of the distribution as 1/n, where n is the number of the data-points

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?

11. Differentiate between Ridge and Lasso Regression.

Ans

|  |  |  |
| --- | --- | --- |
| BASIS | RIDGE | LASSO |
| Technique name | It is also referred as L2 regularization technique | It is also referred as L1 regularization technique |
| Focus work | Ridge will reduce the impact of features that are not important in predicting your y values | Lasso will eliminate many features and reduce overfitting in the linear model |
| Formula | Formula of ridge is addition sum of the error and sum of the squares of coefficient | Formula of lasso is addition of sum of the error and sum of the absolute value of coefficient |
| Slope | Ridge regression can reduce the slope close to zero (but not exactly zero) | Lasso regression can reduce the slope to be exactly equal to zero |
| Parameters | Ridge works well if there are many large parameters of about the same value | Lasso tends to do well if there are a small number of significant parameters and the other are close to zero |

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)- it is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

The suitable value of a VIF for a feature to be included in a regression modelling is three or below three . As VIF increases, the les reliable your regression results are going to be.

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

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

Ans MEAN SQUARED ERROR- the MSE is calculated as the mean or average of the squared difference between predicted and expected target values in a dataset

MEAN ABSOLUTE ERROR-it is a popular metric because, like RMSE, the units of the error score match the unit of the target value that is being predicted. The MAE score is calculated as the average of the absolute error value.

ROOT MEAN SQUARED ERROR- the RMSE is an extension of the mean squared error. The squared root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted. For example if your target variable has the units “dollars”, then the RMSE error score will also have the unit “dollars” and not “squared dollars” like the MSE

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

|  |  |  |
| --- | --- | --- |
| Actual/predicted | True | False |
| True | 1000 | 50 |
| False | 250 | 1200 |

Ans Accuracy = TP+TN/TP+TN+FP+FN

= 0.88

Specification= TN/TN+FP

=0.96

Precision= TP/TP+FP

=0.95

Sensitivity = TP/TP+FN

= 0.8