

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

Answer= 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?

- 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.

Answer= Decision trees 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

Answer= Decision trees

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.

Answer= Precision

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

Answer= MSE

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

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

Answer= 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

Answer= Pruning

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

Answer= 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?

11. Differentiate between Ridge and Lasso Regression.

Answer= lasso and Ridge are regularization techniques

LASSO:

Lasso will try to remove or omit any input variable which is not making effect on output variable

RIDGE:

In any dataset if there is very high difference between the positive and negative correlation ridge regression will try to minimize the difference between those two inputs

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

Answer=The Variance inflation factor (VIF) measures the severity of multicollinearity in regression analysis.

After calculation of VIF we can remove higher vif factor one by one after that my data will be filtered.

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

Answer=Scaling of data come under the source of datasets pre processing,when we are performing machine learning algorithm in data sets.as we know most of the supervised and unsupervised methods are make decision according to the datasets apply to them and algorithm calculate the distance between the data points to make better inference out of the data

For example= if I go a market for purchasing vegetables in shop have various types of vagitables in a bunch of vagetables wev can pick a good one because we have learned about the attributes of vagitables and we know which is good or which is not good,if in that shop all the vagies are at same or good quality we take less time to choose it ,the moral of the example if the values of attributes are closer we will work faster and the chance of getting good vegies are also strong

MACHINE LEARNING

ASSIGNMENT - 6

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