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.

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?

A) SVM

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

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?

A) Adaboost

D) Xgboost.

8. Which of the techniques are used for regularization of Decision Trees?

A) Pruning

C) 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

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?

A10. The basic idea behind the adjusted R2 is to penalize the score as we keep adding new features to the model.

$$Adjusted\ R2 = 1 - (1 - R^2)\frac{(n-1)}{(n-m-1)}$$

Where,
n refers to the number of data points (rows)
m refers to the number of independent features

The denominator (n-m-1) decreases as we increase the value of m. So, if we don’t find a significant increase in R2, then the value of the whole expression will not increase or it may even decrease.

11. Differentiate between Ridge and Lasso Regression.

A11. Lasso Regression (L1), penalizes the sum of the absolute values of the coefficients. This makes Lasso zero out some coefficients

Ridge Regression (L2), the penalty function is defined by the sum of the squares of the coefficients. This makes Ridge prevent the coefficients of your Beta vector to reach extreme values

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

A12. VIF stands for Variance Corelation Factor

VIF measures the strength of the correlation between the independent variables in regression analysis. This correlation is known as multicollinearity.

Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of 5 or even 2.5.

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

A13. Pre-processed data may contain attributes having values of different units or scales. If not scaled the feature with a higher value range will start dominating the output and we may not get desired results. Hence data scaling techniques are used.

Examples: Normalisation and Standardisation

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

A14.Metrics used to check the goodness of fit in linear regression are:

- MAE: Mean Absolute Error
- RMSE: Root Mean Square Error
- R2
- Adjusted R2

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

A15.

- Accuracy: 2200/2500 = 0.88
- Recall/Sensitivity: 1000/1250 = 0.80
- Precision: 1000/1050 = 0.95
- Specificity: 1200/1250 = 0.96