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: 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?
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) 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
ANS: 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 B) Sensitivity C) Precision D) None of the above.
ANS: 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?
A) Model A B) Model B C) both are performing equal D) Data Insufficient
ANS: 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 B) R-squared C) MSE D) Lasso
ANS: A) Ridge & D) Lasso.

7. Which of the following is not an example of boosting technique?
A) Adaboost B) Decision Tree C) Random Forest D) Xgboost.
ANS: B) Decision Tree C) 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: D) All of the above

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:

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:

11. Differentiate between Ridge and Lasso Regression.
ANS: Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros. During training, the objective function become:

$$\frac{1}{2m}\sum_{i=1}^{m}(y-Xw)^2 + alpha\sum_{j=1}^{p}|w_j|$$

As you see, Lasso introduced a new hyperparameter, *alpha*, the coefficient to penalize weights.

Ridge takes a step further and penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. The objective function becomes:

$$\sum_{i=1}^{n}(y-Xw)^2 + alpha\sum_{j=1}^{p}w_j{}^2$$

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
ANS: A variance inflation factor (VIF) 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. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

13. Why do we need to scale the data before feeding it to the train the model?
ANS: Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. As we know most of the supervised and unsupervised learning methods make decisions according to the data sets applied to them and often the algorithms calculate the distance between the data points to make better inferences out of the data.

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

ANS: Regression is a type of Machine learning which helps in finding the relationship between independent and dependent variable.

In simple words, Regression can be defined as a Machine learning problem where we have to predict discrete values like price, Rating, Fees, etc.

Metrics which are use to check the goodness of fit in linear regression:

- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error
- Root Mean Squared Log Error
- R Squared
- Adjusted R Squared

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

ANS: accuracy= 0.8

Precision= 0.8

Recall=0.95

THE END