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?
   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

Correct option - C

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.

Correct option - B

3. Which of the following is an ensemble technique?
   A) SVM
   B) Logistic Regression
   C) Random Forest
   D) Decision tree

Correct option - D

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.

Correct option – B

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

Correct option – 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

Correct option – A, D

MACHINE LEARNING

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

Correct option – B, C

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

Correct option – A, C

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

Correct option – A, B

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?

    The adjusted R-squared make use of the degree of freedom to penalize the unnecessary predictions when the number of independent variable increases then the value of Adjusted R-squared decreases.

11. Differentiate between Ridge and Lasso Regression.
    Ridge regression also called Tikhonov regularization or L2 regularization and it simply add regularization term to the cost function which keeps model weights small as possible, where model weights depends on hyperparameter alpha.

    Least Absolute Shrinkage and Selection Operator(Lasso) Regression, also known as L1 regularization.In Lasso regression it adds L1 norm of weight vector to the cost function which allows Lasso regression to eliminate least important features.

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

    VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. VIF exceeding 5 or 10 indicates high multicollinearity which is needed to include in a regression model.

MACHINE LEARNING

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

The range of all features should be normalized so that each feature contributes approximately proportionately to the final distance and also to ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features.

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

- R Squared
- Adjusted R Squared
- F Statistics
- RMSE / MSE / MAE

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

- Sensitivity/ recall =1000/ (1000+50) =0.95
- Specificity = 1200 / (1200+250) = 0.96
- Precision = 1000 / (1000 + 250) = 0.80
- Accuracy= (1000 + 1200) / (1000 + 50 + 250 + 1200)= 0.88