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

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.

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

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.

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 is the regularization technique in Linear Regression??
A) Ridge
B) R-squared
C) MSE
D) Lasso

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

A) Adaboost

B) Decision Tree

C) Random Forest

D) Xgboost.

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

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

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:- The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability.Adjusted R-squared increases only when independent variable is significant and affects dependent variable.

11. Differentiate between Ridge and Lasso Regression.

Ans:- Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square.

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

Ans: - Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

A VIF of three or below is not a cause for concern. As VIF increases, the less 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?

Ans: - Scaling of the data makes it easy for a model to learn and understand the problem.

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

Ans: - Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE).

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans: -

True Positive= 1000	**True Negative= 50**
False Positive= 250	**False Negative= 1200**

$$\textbf{ACCURACY=} \quad \frac{TP+TN}{TP+TN+FP+FN} = \frac{1000+50}{1000+50+250+1200} = \frac{1050}{2500} = 0.42$$

$$\textbf{RECALL=} \quad \frac{TP}{TP+FN} = \frac{1000}{1000+1200} = \frac{1000}{2200} = 0.45$$

$$\textbf{PRECISION=} \quad \frac{TP}{TP+FP} = \frac{1000}{1000+250} = \frac{1000}{1250} = 0.8$$

$$\textbf{SPECIFICITY=} \quad \frac{TN}{TN+FP} = \frac{50}{50+250} = \frac{50}{300} = 0.166$$

$$\textbf{SENSITIVITY=} \quad \frac{TP}{TP+FN} = \frac{1000}{1000+1200} = \frac{1000}{2200} = 0.45$$