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?
B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?
C) Random Forest

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?
C) 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
D) Lasso

7. Which of the following is not an example of boosting technique?
C) Random Forest

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

9. Which of the following statements is true regarding the Adaboost technique?
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?
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**. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

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.Ridge regression is also referred to as **L2 Regularization**.

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

ANS-

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

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

ANS-

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, we scale the data before feeding it to the model.

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

ANS-

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: **Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)**

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

ANS-

1. Precision (**true** positives / **predicted** positives) = TP / TP + FP
2. Sensitivity OR Recall (**true** positives / all **actual** positives) = TP / TP + FN
3. Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP
4. Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN