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 are 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?
11. Differentiate between Ridge and Lasso Regression.
12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
13. Why do we need to scale the data before feeding it to the train the model?

MACHINE LEARNING

14. What are the different metrics which are used to check the goodness of fit in linear regression?
15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

ANSWERS

Answer 10. 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.

Answer 11. The main difference between Lasso and Ridge is the penalty term they use. Ridge uses L2 penalty term which limits the size of the coefficient vector. Lasso uses L1 penalty which imposes sparsity among the coefficients and thus, makes the fitted model more interpretable.

Lasso tends to do well if there are a small number of significant parameters and the others are close to zero (ergo: when only a few predictors actually influence the response). Ridge works well if there are many large parameters of about the same value (ergo: when most predictors impact the response).

Answer 12. 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.
As a rule of thumb, 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.

Answer 13. 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.

Answer 14. 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)

Answer 15. Calculation is done below:

Accuracy = TP+TN / TP+TN+FP+FN = 2200/2500 = 0.88

Precision = TP / TP+FP =1200/1050 = 0.95

Recall = TP / TP+FN = 1000/1250 = 0.8

Sensitivity = TP / TP + FN = 1000/1250 = 0.8

Specificity = TN / TN + FP = 1200/1250 = 0.96