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 ofthe 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

MACHINE LEARNING

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 themodel?

Answer-

Adjusted R Squared= 1 — [((1 — R Squared) * (n-1)) / (n-p-1)]

whenever the number of independent variables gets increases, it will penalize the formula so that the total value will come down. It is least affected by the increase of independent variables. Hence, Adjusted R Squared will more accurately indicate the performance of the model than the R Squared.

11. Differentiate between Ridge and Lasso Regression.

Answer-

LASSO-This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression.

RIDGE-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 regressionmodelling?

Answer- VIF measures the strength of the correlation between the independent variables in regression analysis. This correlation is known as multicollinearity, which can cause problems for regression models. An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

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

Answer- We scale the data before feeding it to the model 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?

Answer- There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square
2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
3. Mean Absolute Error(MAE)

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

Answer-

MACHINE LEARNING

Accuracy =

1000 + 1200/ 1000 + 1200 + 50 + 250=0.88

Precision =

1000 / 1000 + 50=0.95

Sensitivity / Recall =

1000/ 1000 + 250=0.8

Specificity =

1200 / 1250=0.96

Actual/Predicted	True	False
True	1000tp	50fp
False	250fn	1200tn