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
Ans: A

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.
Ans: B

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

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.
Ans: 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
Ans: 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
Ans: A & D

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

MACHINE LEARNING

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
Ans: D

9. Which of the following statements is true regarding the Adaboost techniquee?
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
Ans: 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?

Ans: The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model. Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

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: A variance inflation factor (VIF) is a measure of the amount of multi collinearity in regression analysis. Multi collinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. The default VIF cutoff value is 5; only variables with a VIF less than 5 will be included in the model. However, note that many sources say that a VIF of less than 10 is acceptable.

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)

MACHINE LEARNING

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans:

1. TP = 1000
2. TN = 1200
3. FP = 50
4. FN = 250

Sensitivity* = *TP/(TP+FN)* = *1000/(1000+250)* =*0.8

Specificity* = *TN/(TN+FP)* = *1200/(1200+50)* = *0.96

Precision* = *TP/(TP+FP)* = *1000/(1000+50)* = *0.95

Recall* = *TP/(TP+FN)* = *1000/(1000+250)* =*0.8

Accuracy* = *(TP+TN)/(TP+FP+FN+TN)* = *(1000+1200)/(1000+50+250+1200)
= 0.88