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

Answer is C. High R-squared value for train-set and Low R-squared value for test-set.

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.

Answer is B. Decision trees are highly prone to overfitting

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

Answer is 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?
   A) Accuracy
   B) Sensitivity
   C) Precision
   D) None of the above.

Answer is A. Accuracy

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

Answer is Model 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

Answer is Ridge & Lasso

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

MACHINE LEARNING

Answer is Random Forest & Decision tree.

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

Answer is Pruning & Restricting the max depth of the tree

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

Answer is below.

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

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?

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. The R-squared, also called the coefficient of determination, is used to explain the degree to which input variables (predictor variables) explain the variation of output variables (predicted variables). It ranges from 0 to 1

11. Differentiate between Ridge and Lasso Regression.

Ridge Regression:
Ridge regression is a technique used to analyze multi-linear regression (multicollinear), also known as L2 regularization. It is Applied when predicted values are greater than the observed values.

Lasso Regression:
Lasso Regression, It is a technique where data points are shrunk towards a central point, like the mean. Lasso is also known as L1 regularization.
It is applied when the model is overfitted or facing computational challenges.

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

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.
A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of

MACHINE LEARNING

the variable with other variables.
In general, if it is less than 10, it is ok. However, We may check the data, which factor has high correlation with that factor. If an outlier exists? I.e we need to do some analyses before make decision.

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

Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem.
scaling of the data makes it easy for a model to learn and understand the problem, an independent variable with a spread of values may result in a large loss in training and testing and cause the learning process to be unstable.
Normalization and Standardization are the two main methods for the scaling of the data. Both of them can be implemented by the scikit-learn libraries preprocess package.

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

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship.

different metrics are below we use in regression.

MSE, MAE, R-squared, Adjusted R-squared, and RMSE.

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

Actual/Predicted	True	False
True	1000 (TP)	50 (FN)
False	250 (FP)	1200 (TN)

Accuracy – 0.88

Sensitivity – 0.95

Specificity - 0.82

Precision - 0.80

Recall - 0.95