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 : 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 : 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

Answer : 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 : 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

Answer : B) 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 : A) Ridge D)Lasso

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

Answer : B) Decision Tree C) Random Forest

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 : A) Pruning

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 : d) None of the above

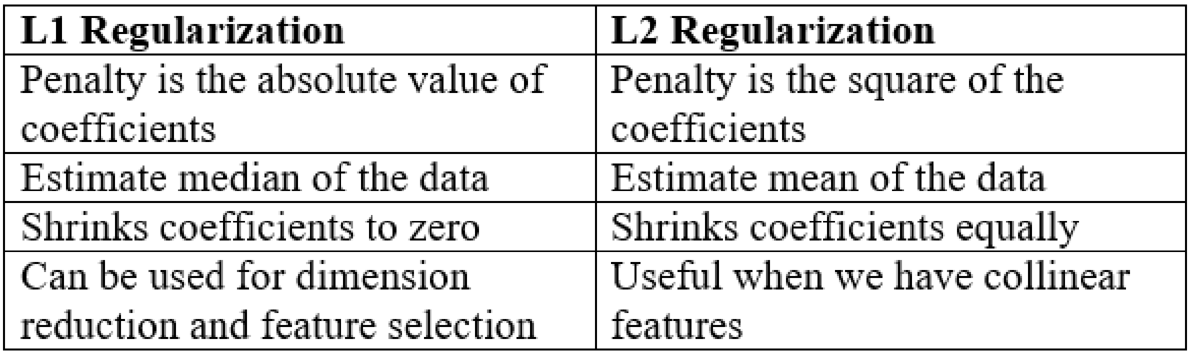
Q10 to Q15 are subjective answer type questions, Answer them briefly.

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

Answer : Adjusted R-squared is **a modified version of R-squared that has been adjusted for the number of predictors in the model**. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

11. Differentiate between Ridge and Lasso Regression.
Answer :

Lasso Ridge



L1 Regularization	L2 Regularization
Penalty is the absolute value of coefficients	Penalty is the square of the coefficients
Estimate median of the data	Estimate mean of the data
Shrinks coefficients to zero	Shrinks coefficients equally
Can be used for dimension reduction and feature selection	Useful when we have collinear features

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

Answer : Variance Inflation Factor is method of finding out multicollinearity in a data set. Any value up to 5 is acceptable for including a feature in regression modeling.

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

Answer : If the data is not scaled it may affect the model badly as on data point may be very small and another may be very high so the learning will not be good and machine will not perform well while testing.

MACHINE LEARNING

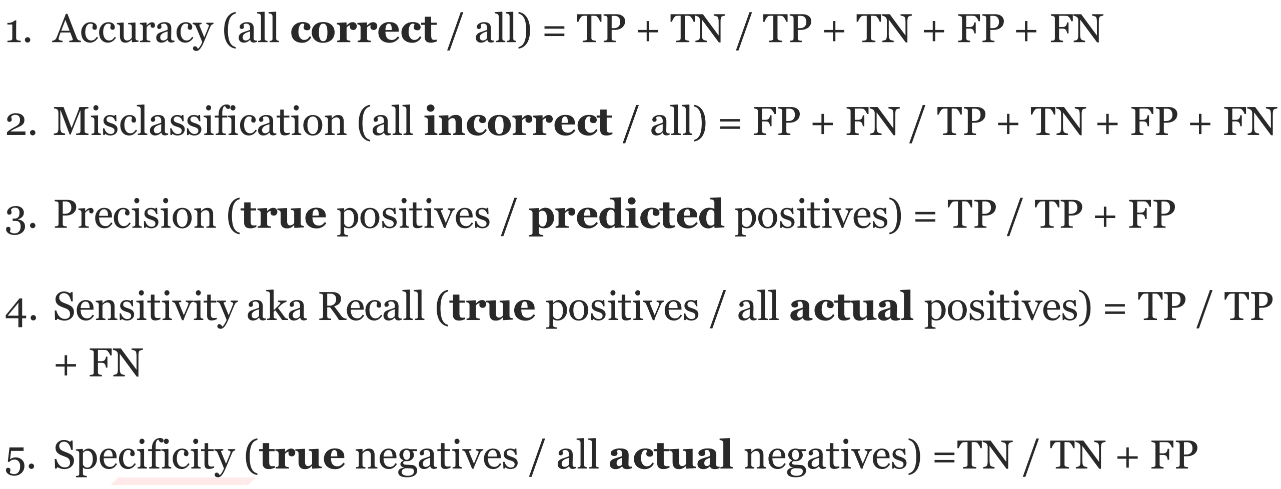
14. What are the different metrics which are used to check the goodness of fit in linear regression?

    Answer : r2 score, score, root mean squared error, mean squared error

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

Actual/Predicted	True	False
True	1000 TP	5 FN
False	250 FP	1200 TN

accuracy = (1000+1200)/2500 = 0.88
precision= (1000/1000+250) = 0.8
recall = sensitivity = 1000/1005 = 0.99
specificity = 1200/1200+250 =82.7



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