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?

Adjusted R² is a modified version of R² adjusted with the number of predictors. It penalizes for adding unnecessary features and allows a comparison of regression models with a different number of predictors.

MACHINE LEARNING

$$Adjusted\ R^2 = \bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Here ***k*** is the number of explanatory variables in the model and ***n*** is the number of observations.

11. Differentiate between Ridge and Lasso Regression.

The main difference between Ridge and LASSO Regression is that **if ridge regression can shrink the coefficient close to 0** so that all predictor variables are retained. Whereas LASSO can shrink the coefficient to exactly 0 so that LASSO can select and discard the predictor variables that have the right coefficient of 0.

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

- Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
- Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.
- VIF measures the number of inflated variances caused by multicollinearity.
- Suitable VIF value is less than 5.

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

Scaling is better to be done in general, because if all the features are on the same scale, the Gradient Descent Algorithm converges faster to the global or optimum local minimum.
We can speed up gradient descent by having each of our input values in roughly the same range. This is because our model parameters, will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

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

- MAE Mean Absolute Error
- MSE Mean Squared Error
- RMSE Root Mean Squared Error

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Sensitivity or Recall – 95.2%
Specificity – 3.44%
Precision – 80%
Accuracy – 88%