MACHINE LEARNING

Q1 to Q5, only one correct answer, 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: 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: 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. Precision
c. Sensitivity
d. None of the above

Answer: c. Sensitivity

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 **and** 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 **and** 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 **and** c. 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: 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?**

    Answer: The adjusted R-squared is a modified version of the R-squared that penalizes the addition of unnecessary predictors to the model. As more predictors are added to the model, the R-squared value will typically increase, even if the predictors are not useful. The adjusted R-squared accounts for this by adjusting the R-squared value based on the number of predictors in the model. It is a more reliable indicator of the goodness of fit of a model and is useful in comparing models with different numbers of predictors.

11. Differentiate between Ridge and Lasso Regression?

Answer: Ridge and Lasso regression are both regularization techniques used to prevent overfitting in linear regression. Ridge regression adds a penalty term to the cost function that is proportional to the square of the magnitude of the coefficients. This helps to shrink the coefficients of less important predictors towards zero, but does not set any coefficients exactly to zero. Lasso regression, on the other hand, adds a penalty term to the cost function that is proportional to the absolute value of the coefficients. This can result in some coefficients being set exactly to zero, effectively performing feature selection.

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

Answer: VIF stands for Variance Inflation Factor, it is a measure of how much the variance of the estimated regression coefficients are increased because of collinearity. A high VIF indicates that the corresponding predictor is highly correlated with one or more of the other predictors. A VIF value of 1 indicates that there is no correlation between this predictor and any other predictors, while a value greater than 1 indicates that there is correlation. A suitable value of VIF for a feature to be included in a regression modeling is typically less than 5 or 10.

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

Answer: Scaling the data is important before training a model because many machine learning algorithms use distance based calculations. So if the data is not scaled, then the algorithm will be sensitive to the scale of the data. For example, if one feature is measured in kilometers and another feature is measured in meters, then the algorithm will be biased towards the feature measured in kilometers. Scaling the data ensures that all the features are on the same scale, which leads to a fair comparison of the importance of each feature.

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

Answer: There are several metrics used to check the goodness of fit in linear regression, some examples are:

R-squared: R-squared measures the proportion of variation in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, where 1 indicates a perfect fit.

Mean Squared Error (MSE): MSE is the average of the square of the residuals, it measures the average difference between the predicted values and the true values.

Root Mean Squared Error (RMSE): it is the square root of MSE and it gives the error in the same unit as the response variable.

Mean Absolute Error (MAE): it is the mean of the absolute values of the residuals.

Adjusted R-squared: It is a modified version of the R-squared that penalizes the addition of unnecessary predictors to the model.

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

Answer: Sensitivity (also known as recall) = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8

Specificity = True negatives / (True negatives + False positives) = 1200 / (1200 + 50) = 0.96
Precision = True Positives / (True Positives + False Positives) = 1000 / (1000 + 50) = 0.95
Recall = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8
Accuracy = (True Positives + True negatives) / (Total) = (1000 + 1200) / (1000+50+250+1200) = 0.89
Note: sensitivity and recall are the same.