MACHINE LEARNING

ASSIGNMENT - 6

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:- 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.
Ans:- 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
Ans:- 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.
Ans:- 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
Ans:- 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
Ans:- A) Ridge or D) Lasso

7. Which of the following is not an example of boosting technique?
A) Adaboost B) Decision Tree
C) Random Forest D) Xgboost.
Ans:- B) Decision Tree or 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
Ans:- 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
Ans:- A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points OR
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?
Ans:- Use predicted R-squared to determine how well a regression model makes predictions. This statistic helps you identify cases where the model provides a good fit for the existing data but isn't as good at making predictions. However, even if you aren't using your model to make predictions, predicted R-squared still offers valuable insights about your model.

- It removes a data point from the dataset.
- Calculates the regression equation.
- Evaluates how well the model predicts the missing observation.
- And repeats this for all data points in the dataset.

Predicted R-squared helps you determine whether you are overfitting a regression model. Again, an overfit model includes an excessive number of terms, and it begins to fit the random noise in your sample.

By its very definition, it is not possible to predict random noise. Consequently, if your model fits a lot of random noise, the predicted R-squared value must fall. A predicted R-squared that is distinctly smaller than R-squared is a warning sign that you are overfitting the model. Try reducing the number of terms.

11. Differentiate between Ridge and Lasso Regression.
Ans:- Ridge Regression :- A regression model that uses L2 regularization technique is called Ridge regression. It adds "squared magnitude of coefficient as penalty term to the loss function
Lasso Regression :- A regression model which uses L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression. It adds "obsolute value of magnitude of coefficient os penalty term to the loss function

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

Ans:- VIF :- 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 above 10 indicates high correlation and is cause for concern. Some authors suggest the suitable value of a VIF for a feature to be included in a regression modelling more conservative level of VIF is 2.5 or above.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

Ans:- Scaling the value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem in the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.

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)

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Actual/Predicted	Predicted True	Predicted False
Actual True	1000	50
Actual False	250	1200

When TP = 1000, TN = 1200, FP =250, FN = 50

1. Accuracy = $TP + TN /(TP + TN + FP + FN)$

   1000 + 1200 / (1000 + 1200 + 50 + 250)

   = .88 Its means = 88%

2. Sensitivity/ Recall = $TP / TP + FN$

   = 1000/ 1000 + 50

   = .95

3. Precision = $TP / TP + FP$

   = 1000/ 1000 + 250

   = 0.8

4. Specificity = $TN / TN + FP$

   = 1200 / 1200 + 250

   = .82

5.
6.