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-

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,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, 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- 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

ANS- 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- The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance

11. Differentiate between Ridge and Lasso Regression.

ANS-Ridge and lasso regression are two common machine learning approaches for constraining model parameters. Both methods try to get the coefficient estimates as close to zero as possible because minimizing (or shrinking) coefficients can reduce variance dramatically (i.e., overfitting)

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

ANS- The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

As a rule of thumb, a VIF of three or below is not a cause for concern. As VIF increases, the less reliable your regression results are going to be.

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

ANS- To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to 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 comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set.

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

ANS-So basically R2 squared calculates how must regression line is better than a mean line. Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

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.

ANS-

Actual/Predicted	True	False
True	1000	50
False	250	1200

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

Sensitivity=TP / TP + FN
=0.8

Specificity= TN / TN + FP
=0.96

$$Precision = \frac{TP}{TP + FP}$$

Precision=0.952

$$Recall = \frac{TP}{TP + FN}$$

Recall=0.8

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy=0.88