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

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?

Answer:-- Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R^2 tends to optimistically estimate the fit of the linear regression.

Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11. Differentiate between Ridge and Lasso Regression.

Answer:--

1). In Ridge regression, we add a penalty term which is equal to the square of the coefficient. Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function.

2). Lasso tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

3). Ridge regression decreases the complexity of a model but does not reduce the number of variables since it never leads to a coefficient been zero rather only minimizes it. If the number of predictors *(p)* is greater than the number of observations *(n)*, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant (or may be used in the test set).

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

Answer:---VIF-- 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.

VIF=   1   =   1

___1-Ri^2   Tolerance

Where $\mathbf{R}_i^2$ represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones. The reciprocal of VIF is known as **tolerance**. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.

a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

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

Answer:-- 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.

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.

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

Answer:--- 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).

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data—how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

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

Actual/predicted	True	False
True	1000	50
False	250	1200

Answer:---

Sensitivity formula=TP/(TP+FN)

1000/(1000+250)=.8

Specificity formula=TN/(TN+FP)

1200/(1200+50)=.96

Precision=TP/(TP+FP)

1000/(1000+50) = .95

Recall = TP/(TP+FN)

1000/(1000+250) = .8

Accuracy = TP+TN/TP+TN+FP+FN

1000+1200/1000+1200+50+250 = .88