MACHINE LEARNING Assigment -6

Q1. 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

➔C

Q2. 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

➔B

Q3. Which of the following is an ensemble technique?
A) SVM B) Logistic Regression
C) Random Forest D) Decision tree

➔C

Q4. 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.

➔B

Q5. 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

➔B

In Q6 to Q9, more than one options are correct, Choose all the correct options:
Q6. Which of the following are the regularization technique in Linear Regression??

A) Ridge B) R-squared
C) MSE D) Lasso

➔A,D

Q7. Which of the following is not an example of boosting technique?
A) Adaboost B) Decision Tree
C) Random Forest D) Xgboost.

➔B,C

Q8.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

➔A,C

Q9. 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

➔A,B

Q10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?
➔

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R Squared has no relation to express the effect of a bad or least significant independent variable on the regression.
Compared to R Squared which can only increase, Adjusted R Squared has the capability to decrease with the addition of less significant variables, thus resulting in a more reliable and accurate evaluation. Adjusted R-2 is able to penalize as it considers the degree of freedom factors. Degree of freedom is given by:
d.o.f. = n-k-1
where,
k is no if independed varialbed
n is the number of observation

Adjusted R Squared, however, makes use of the degree of freedom to compensate and penalize for the inclusion of a bad variable.
Adjusted R Squared can be expressed as :
Adj R2 = 1 – (1-R2) * (n-1)/d.o.f.
That is , Adj R2 = 1 – (1-R2) * (n-1)/n-k-1

The value of Adjusted R Squared decreases as k increases also while considering R Squared acting a penalization factor for a bad variable and rewarding factor for a good or significant variable. Adjusted R Squared is thus a better model evaluator and can correlate the variables more efficiently than R Squared.

Q11.Differentiate between Ridge and Lasso Regression.

➔ In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The cost function of ridge function looks like:
Cost function= MSE+ α * (sum of square of coefficients)2

Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from 0 this term penalizes, cause model, to decrease the value of coefficients in order to reduce loss.

Cost function= MSE+ α * |sum of magnitude of coefficients|

The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

Q12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

➔ A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIF = 1 / (1 – R^2)

VIF of 2.5 or above but less than 9 is suitable for regression modeling.

Q13. Why do we need to scale the data before feeding it to the train the model?

➔

Scaling is one of the important pre-processing that is required for standardizing/normalization of the input data. When the range of values are very distinct in each column, we need to scale them to the common level. The values are brought to common level and then we can apply further machine learning algorithm to the input data. One way to scale the values is to bring the values of all the column between 0 to 1 or we can bring them to common level having values between -3 to 3.
The other is gradient descent:
The gradient descent algorithm which is used to reach the optimal solution in most of the cases, it reached the optimal solution much faster if all the features are at the same scale. That's why scaling helps to reach the optimal solution.

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

➔The metrics that we can use to check the goodness of fit for linear regressions are as follow:Mean Absolute Error and Mean Square Error.Root Mean Squared Error,Relative Absolute Error and Relative Squared Error

R^2 and Adjusted R^2

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: 0.9523

SPECIFICITY: 0.8275

ACCURACY: 0.88