## Submitted by: Monika Singh

# Machine Learning Worksheet - 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 notperforming 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 themodel?

Ans:

R-squared (R2) is a statistical measure that shows how much of a dependent variable's variance is explained by one or more independent variables in a regression model. No relationship exists between R Squared and the way a poor or least significant independent variable affects the regression. Modified R Squared has the capacity to decrease when less important factors are added, in contrast to R Squared which can only grow. This makes it a more dependable and accurate evaluation method. As it takes into account the degree of freedom elements, adjusted R-2 has the ability to penalize. Degree of freedom is given by:

Degree of freedom = n-k-1

Where, k = no if independed variable

n = 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 is given by:

Adj R2 = 1 – (1-R2) * (n-1)/ Degree of freedom

That is, Adj R2 = 1 – (1-R2) * (n-1)/n-k-1

By taking into account that R Squared acts as a rewarding factor for a good or major variable and a penalizing factor for a bad or insignificant variable, the value of Adjusted R Squared declines as k grows. As a result, adjusted R squared is a superior model evaluator and correlates the variables more effectively.

11. Differentiate between Ridge and Lasso Regression.

Ans:

Ridge and Lasso regression uses two different penalty functions for regularization. Ridge regression uses L2 on the other hand lasso regression go uses L1 regularization technique. In ridge regression, the penalty is equal to the sum of the squares of the coefficients and in the Lasso, penalty is considered to be the sum of the absolute values of the coefficients. In lasso regression, it is the shrinkage towards zero using an absolute value (L1 penalty) rather than a sum of squares (L2 penalty).

As we are aware that the coefficients in ridge regression cannot be 0. The Lasso regression algorithm technique, however, conducts both parameter shrinkage and feature selection concurrently and automatically since it nulls out the coefficients of collinear features. Thus, we either consider all the coefficients or none of the coefficients. This makes choosing a variable or variables from the set of n variables while performing lasso regression simpler and more precise.

The cost function for both ridge and lasso regression are similar. However, ridge regression takes the square of the coefficients and lasso takes the magnitude.

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

Ans:

Regression analysis can identify multicollinearity using a variance inflation factor (VIF). Multicollinearity occurs when there is correlation among predictors in a model; it can have a negative impact on the outcomes of your regression analysis. The VIF calculates the amount of multicollinearity in the model that causes the variance of a regression coefficient to be overstated.

VIF = 1 / (1 – R^2)

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

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

Ans :

One crucial pre-processing step needed to normalize and standardize the incoming data is scaling. We must scale the values to the common level when each column's range of values is significantly distinct. After the values are levelled off, we can continue applying machine learning algorithms 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.

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

Ans:

The metrics that we can use to check the goodness of fit for linear regressions are as follow:

A Mean Absolute Error and Mean Square Error
B Root Mean Squared Error
C Relative Absolute Error and Relative Squared Error
D 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 |

Ans:

SENSITIVITY OR RECALL: 0.9523
SPECIFICITY: 0.8275
ACCURACY: 0.88