Machine learning assignment Answers

Q1 – 5

1. B) Low R-squared value for train-set and high R-squared value for test.

2. C) Decision tees are not easy to interpret

3. C) Random forest

4. A) Accuracy

5. B) Model B

Q6 – 9

6. D) Lasso , A) Ridge

7. A) Adaboost D) Xgboost

8. D) all of the above

9. B) and C)

Q10 – 15

10. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.

11. This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. In this shrinkage technique, the coefficients determined in the linear model are shrunk towards the central point as the mean by introducing a penalization factor called the alpha α (or sometimes lamda) values. Alpha (α) is the penalty term that denotes the amount of shrinkage. Therefore, lasso regression shrinks the coefficients and helps to reduce the model complexity and multi-collinearity.

Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square. Ridge regression is also referred to as L2 Regularization.

12. A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

13. 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. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.

14. 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)