Machine Learning Assignment-6

Answers:

1. C) High R-squared value for train-set and Low R-squared value for test-set.
2. B) Decision trees are highly prone to overfitting.
3. C) Random Forest
4. B) Sensitivity
5. B) Model B
6. A) Ridge, A) Ridge
7. B) Decision Tree, C) Random Forest
8. A) Pruning, C) Restricting the max depth of the tree
9. 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

10. The adjusted R 2 "penalizes" you for adding the extra predictor variables that don't improve the existing model. It can be helpful in model selection. Adjusted R 2 will equal R 2 for one predictor variable. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. **Ridge Regression:** It is used to solve multi collinearity in OLS regression models through the incorporation of shrinkage parameter (it is vital in ridge regression).The assumptions for the model is same as OLS model like linearity, constant variance and independence and normality not need to be assumed.
    Lasso Regression: It is more similar to Ridge Regression but perform automatic variable selection. It allows regression coefficient to be zero whereas Ridge does not.

12. 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-R2i
    Most statistical software has the ability to compute VIF for a regression model. The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows: A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.

13. 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**. In the case of neural networks, an independent variable with a spread of values may result in a large loss in training and testing and cause the learning process to be unstable. Scaling gives equal weights/importance to each variable so that no single variable steers model performance in one direction just because they are bigger numbers.

14. There are three main metrics which are used to check the goodness of fit in linear regression.

1. **R Square/Adjusted R Square:** R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE): While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

3. Mean Absolute Error(MAE): Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

15.

Actual/Predicted	True	False
True	1000 TP	50 FP
False	250 FN	1200 TN

Accuracy = TP+TN/TP+TN+FP+FN = 1000+1200/1000+1200+250+50 =2200/2500 = 0.88
Recall = TP/TP+FN = 1000/1000+50 = 0.95
Precision = TP/TP+FP = 1000/1000+250 = 0.8
Sensitivity = TP/TP+FN = 1000/1000+50 = 0.95
Specificity = TN/FP+TN = 1200/50+1200 = 1200/1250 = 0.96