

MACHINE LEARNING

Worksheet set 6 (Answers)

1. High R-squared value for train-set and Low R-squared value for test-set.
2. Decision trees are highly prone to overfitting.
3. SVM
4. Precision
5. Model B
6. Ridge and Lasso
7. Decision Tree and Random Forest
8. Pruning and Restricting the max depth of the tree
9. A tree in the ensemble focuses more on the data points on which the previous tree was not performing well and None of the above
10. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhance the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.
11. Lasso and 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. Lasso regression is also referred to as L1 regularization and Ridge regression is referred as L2 regularization.
12. VIF stands for Variance inflation factor, VIF is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. suitable value of a VIF for a feature to be included in regression is to less than 5.
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 train the model.
14. Three statistics are used to check the goodness of fit in linear regression
 - i) R-squared
 - ii) overall F-test
 - iii) Root Mean Square Error.
15. Sensitivity = 0.8, specificity = 0.9, precision = 0.95, recall= 0.12, accuracy = 0.88