Department of Mechanical Engineering

ME 781: Engineering Data Mining and Applications

Assignment-2

1. Read in the data set **a2-data-set.csv** into a data frame.
2. Partition the data set as follows:
    a) Training data: D1: First 400 observations
    b) Testing data: D2: Next 100 observations
3. Using the Regression Tree method and the data set D1 do the following:
    a) Tree1 => Create a tree by mandating at least 5 observations per leaf node
    b) Tree2 => Create a tree by mandating at least 10 observations per leaf node
    c) Tree3 => Create a tree by mandating at least 20 observations per leaf node
    d) Tree4 => Create a tree by mandating at least 40 observations per leaf node
4. Read in data set **a2-data-set-D3.csv** into a data frame D3: this is a sample derived from the population
5. Read in data set **a2-data-set-D4.csv** into a data frame D4: this is another sample derived from the population
6. Now calculate the following for <u>**each**</u> of the above trees:
    a) Residuals with respect to D1, D2, D3 and D4: Name them as RD1, RD2, RD3 and RD4
    b) RMSE and MAE with respect to D1, D2, D3, D4: Name them as RMSE_D1, MAE_D1, and so on.
    c) Capture the data generated so far in Table-1 as follows:

	Tree1	Tree2	Tree3	Tree4
RMSE_D1	RMSE_D1	RMSE_D1	RMSE_D1	RMSE_D1
RMSE_D2				
RMSE_D3				
RMSE_D4				
MAE_D1	MAE_D1	MAE_D1	MAE_D1	MAE_D1
MAE_D2				
MAE_D3				
MAE_D4				

7. Based on information captured in Table-1 evaluate and comment on the models in terms of the following:
    a) With respect to the training data (D1): What trend do you observe in RMSE and MAE when leaf node size increases from 5 to 40 observations. Why is this trend observed?
    b) When you compare performances on D1 v/s D2, in which case is it better? Why?
    c) For a given tree, say Tree3, when you compare it’s performance across D1 … D4
        i. What do you observe? Do you see any effect of Bias-Variance trade-off across these cases?

        ii. What can you do to reduce both Bias and Variance?

8. For the same data sets D1 … D4 create additional tree-based models as follows:
    a) Tree5 => Use the BAGGING method, with at least 20 observations per leaf node and 100 trees
    b) Tree6 => Use RANDOM FOREST method, with at least 20 observations per leaf node, and 100 trees
    c) Tree7 => Use ADABOOST method, with at least 20 observations per leaf node (if possible), and 100 trees
    d) Again, capture the data in Table-2 as follows:

	Tree5	Tree6	Tree7
RMSE_D1	RMSE_D1	RMSE_D1	RMSE_D1
RMSE_D2			
RMSE_D3			
RMSE_D4			
MAE_D1	MAE_D1	MAE_D1	MAE_D1
MAE_D2			
MAE_D3			
MAE_D4			

9. In the context of Tree5 … Tree7:
    a) How do you rate their performances? Why do you see these differences?
    b) How do these models compare with Tree1 … Tree4? Why do you see these differences?
    c) What do you have to say about the Bias/Variance characteristics of Trees 5-7, when compared amongst themselves?
    d) What do you have to say about the Bias/Variance characteristics of Trees 5-7 when compared with Trees 1-4? Explain the differences, if any.
10. Select a method: say BAGGED tree. By using the training set D1 for building the tree and the testing set D2 for evaluating the tree …
    a) Find out values for the following hyper-parameters for which you obtain the BEST model:
        i. Minimum nodes per leaf
        ii. Number of trees to be used in the BAGGING model

Note:

- Create a document containing all your Tables, graphs and conclusions. Number them based on the questions above.
- Save all your created data sets. Name the data sets D1 …D4 into unique csv files
- Place the document and data sets into a directory named Assignment-2
- Zip up the directory and submit it to the assignment submission point in Moodle

- The Test scheduled on Nov-2-2018 will assume you have completed this assignment.