**ENGR-421 HW-5 FALL-2021**

**Name-Surname: Barış KAPLAN**

**KU ID Number: 69054**

Firstly; I have read the csv file,divided the data set into 2 columns (the first column includes the x values and the second column includes y values), and divided each column into the training data set and test data set. While reading the csv file, I have used np.genfromtxt() function of the numpy library.

Then, for learning a tree, I have written a decision tree algorithm by using the preprunning rule given in the “hw05\_description.pdf” file. In this algorithm, my main aim was to find the best split positions for all the nodes (same logic with Lab07-Decision Trees). In this algorithm, I have calculated the errors by using “sum of squares total” approach for regression.

After that, I have assigned the value of the pre-prunning parameter (which is P) to 25, and I have learned a decision tree. Subsequently, I have drawn the fit & test data points & training data points in a graph. In figure 3, you can see the graph in which I have drawn the fit, the test data points, and the training data points.

Next, I have calculated the RMSE values for the training data set and for the test data set. While calculating the rmse values, I have utilized the formulas in Figure 1 (I have used the first (left) formula for calculating the RMSE for the training data points, and the second (right) formula for calculating the RMSE for the test data points). I have found the RMSE results in the Figure 2.

A picture containing text

Description automatically generated

Text

Description automatically generated

Figure 2: The RMSE results I found

Figure 1: The RMSE formulas I have used for the training set and test set (the first formula for the training set, and the second formula for the test set).

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Then; by assigning the value of the preprunning parameter to the each of the values in the array [5,10,15,20,25,30,35,40,45,50] (step size=5), I have learned the decision trees. I have obtained the RMSE values for the preprunning parameter values in the array [5,10,15,20,25,30,35,40,45,50]. Subsequently; for training and test data points, I have plotted the “PrePrunning Size Values vs RMSE Values” graph. In the Figure 4, you can see the “PrePrunning Size Values vs RMSE Values” graph. In the Figure 4, I have obtained the RMSE values for the test data points and for the training data points.

The array which includes the preprunning parameter values= [5,10,15,20,25,30,35,40,45,50]

Figure 4: The “Preprunning Size Values (P) vs RMSE Values” graph I have found

Figure 3: The Fit (Regressogram) & The Training Data Points & The Test Data Points