ENGR-421 HW5

Berke Can Rizai 69282

We import the libraries,

Graphical user interface, text, application

Description automatically generated

After reading from the csv file provided,

We make train and test split,

Graphical user interface, text, application

Description automatically generated

To implement the tree structure, we use the dictinories from the lab that we have implemented, Graphical user interface, text

Description automatically generated

These are representing our nodes, their values and the decision criteria for each feature (dimension).

I defined a function that takes the left, right indices and the data target values and returns the error for pruning.

In the tree implementation, unlike the lab session, we need to keep it inside a function so that we can test and evaluate for each P value easily. P is the maximum depth.

The method signature is as follows,



Inside the while loop is quite similar to lab implementation, unlike the lab, we don’t have frequency.

I added pruning as well as a few checks for purity inside the whole function a few times as it was not producing good results otherwise. Sometimes it would break early.

In the split\_scores, we hold the scores for each of the split.

Text

Description automatically generated

Here, we sort the values and we don’t add the non unique ones as we want to find a place to split the data. We try them and get the best middle value out all of them. They are sorted because otherwise we wouldn’t find split as it is random.

Further,

Text

Description automatically generated with medium confidence

In the get result function, I return the result by traversing the tree and if the current value is smaller, I go to the left and if it is larger I go to the right node by i\*2 + 1. Since this is array implementation.

We then,

Pget results and print them on the plot

Chart, scatter chart

Description automatically generated

Both for train and test splits.

A screenshot of a computer

Description automatically generated with medium confidence

rmseForP returns the mean squared error for the pruning factor P.

I iterate over the test and train and for each of them I deduct prediction from real value and square.

Graphical user interface, text, application, Word

Description automatically generated

Plist holds values for P from 5 to 55 by 5.

In the ad list, I hold scores for the rmse for each of the Ps.

Then, I plot it with the plt lib.

Chart, line chart

Description automatically generated