ECE5984 SP22 - Prof. Jones – HW6

Due Thursday, April 28, 2022 – 11:59 PM via Canvas

In this assignment you are to develop a two-stage regression model to predict the fare for a taxi ride. This two-stage model will consist of three first-stage models, operating on the dataset and trained on the target variable as usual. These will be followed by a second stage that combines the results of the three first-stage models to produce the final output.

Here is what you are to do.

1. Load the "Taxi_Trip_Data.xlsx" dataset. Use the data on the sheet named "taxi_tripdata". This data has a continuous target called "total_amount".
2. Remove columns that are not useful: "store_and_fwd_flag", "PULocationID", and "DOLocationID".
3. Examine the columns "lpep_pickup_datetime", "lpep_dropoff_datetime", "PUBorough" and "DOBorough". See what (if any) use can be made of these columns to create modeling variables.
4. Since our goal is to predict the fare for a taxi ride before it happens, remove columns that would not be known in advance: "fare_amount", "extra", "mta_tax", "tip_amount" and "tolls_amount".
5. Normalize the data appropriately.
6. Divide the data into training and test sets. Use a specified random seed so the split is done the same way each time.
7. Train three regression models on this data:
    a. A regression neural network;
    b. A regression decision tree; and
    c. A multivariate linear regression module.
    For each of these models, choose appropriate architectures and parameters.
8. Train a second-stage regression model that has as inputs the outputs of these three models. Choose an appropriate model type, architecture and parameters.
9. For each of the four models, report the performance as MSE, MAE, R2 and EVS.
10. For the second-stage model, plot the learning curve (training and validation loss by epoch) and a scatterplot of model output versus actual output for each sample in the test set.
11. Summarize your results and your findings.

Your submission should be a Word or PDF document including a description of your final architecture, including a simple drawing, the performance numbers for all four models, your two graphs, and all of your code (pasted in as unformatted text, not formatted on a dark background or as a screen shot!). Also attach your Python code file(s) as .py files (not as .ipnyb files).
Submit your work in the usual way via Canvas.