ASSIGNMENT 3

4th May , 2023

Task 1

Problem

Find out how well can you predict the O3 and NO2 using the method suggested by the manufacturer. To do this, learn the best linear model that uses just the 4 voltage values to predict O3 and NO2 values. Remember that for this part, you cannot use non-linear models, nor can you use temp, humidity, time stamp as features. However, you can use different loss functions e.g. least squares loss, absolute loss, ϵ-insensitive loss as well as different regularizers e.g. ridge, lasso etc. If you are trying out support vector regression for this part, remember to use the linear kernel.

Describe the method that gave you the best-performing linear model (in terms of MAE on training data) and write down what mean absolute error (MAE) does your model give on the training set.

Solution

We tried various combinations of least square loss and absolute loss with regularizers like Ridge, Lasso and ElasticNet. But the results with and without regularizers were very similar since the linear model under fits the data. Thus our best-performing linear model is a Linear Regression model which is trained using the least squares loss on the given training data using only 4 features i.e. the 4 voltage values. It doesn't perform regularization.

MAE on training data (for O_3) = **5.6259**
MAE on training data (for NO_2) = **6.5401**

Task 2

Problem

Chances are that you may not get a very satisfactory result using just a linear model and just the voltage features. Thus, in this next part, develop a learning method that is free to use temp, humidity, time stamp in addition to the voltage features to predict the O3 and NO2 values. You are also free to use non-linear models e.g. decision trees, kernels, nearest-neighbors, deep-nets, etc. Describe the method you found to work best giving all details of training strategy e.g. choice of loss function and tuning of hyperparameters.

Note that you may or may not find the time stamp as a useful feature since some of these pollutants are known to have a diurnal cycle e.g. Ozone is known to have high values during the daytime when sunlight is abundant and low values during night time due to darkness.

Solution

We transformed the time feature string to an integer by making use of only the hour indicator. This is because exact minutes and seconds do not matter that much. The introduction of temperature and humidity features improved the performance of linear models by 0.3 improvements in MAE, while the time feature also improved MAE by 0.03 units. The introduction of these features still results in the underfitting of training data using linear models. So, we tested KNN, Decision Tree, and Random Forest approach under various settings of hyperparameters which gave much better performance but also got a little overfitted. Due to overfitting, we decided to split the data into train and validation set to get the optimal value of hyperparameters. Of these, Random Forest Regressor with 500 estimators gave the minimum MAE of 3.49 on the validation data. But since it uses 500 decision trees to perform the prediction, the model size rose up to 700 MBs. It is also higher prone to overfitting. So, we present details two models, first, K-Nearest Neighbours Regressor, which gives the second-best MAE but is optimal in size and prediction time, and second, the Random Forest Regressor, which gives the best MAE but has high space requirements.

1. **K-Nearest Neighbours:**
   Loss function and hyperparameters -
   - Number of neighbours: 6 (for O3) and 4 (for NO2)
   - Weights: 'uniform' - All points in each neighborhood are weighted equally
   - Algorithm: Ball Tree Algorithm is used for computing nearest neighbors
   - Leaf Size: 30 (used in Ball Tree algorithm)
   - Metric for distance calculation: Euclidean distance

   Results (obtained on training data = 20k samples) -
   - MAE (for O3): **3.0194**
   - MAE (for NO2): **2.0417**
   - MAE on validation data (O2): 3.7352
   - Prediction Time: 175 ms
   - Model Size (including both models): 3.02 MB
2. **Random Forest:**
   Loss function and hyperparameters -
   - Number of trees: 500
   - Loss function: squared error
   - Max Depth of the tree: None - Nodes are expanded until all leaves are pure or until all leaves contain less than Min Samples Split samples
   - Min Samples Split - minimum number of samples required to split a node: 2
   - Min Samples Leaf - minimum number of samples required to be at a leaf node: 1
   - Max Features - the number of features to consider when looking for the best split - 7 (look at all features for splitting)

   Results (obtained on training data = 20k samples) -
   - MAE (for O3): **1.2415**
   - MAE (for NO2): **0.8123**
   - MAE on validation data (O2): 3.4951
   - Prediction Time: 1250 ms
   - Model Size (including both models): 1.49 GB

Of these, we submitted the first model as it has a low space requirement and gives only a slightly higher MAE (on the validation set) than Random Forests.