CS771A

Assignment-3
Group A-top

1 Problem

Monitoring air quality is of crucial importance for a country like India which is home to some of the most polluted cities in the world. India imports sensors required to measure levels of harmful pollutants like ozone O3 and nitrogen dioxide NO2 but these are usually manufactured in nations with distinct weather conditions like China or European countries so the sensors do not work well right out of the box in Indian conditions.

To get them working, we need to perform a task called *calibration* that looks a lot like regression. In this task, we will calibrate two sensors, one measuring the level of O_3 and another measuring the level of NO_2. Both these sensors are electrochemical in nature i.e. in response to changing levels of the pollutant they are measuring, they output two voltages called OP1 and OP2. More specifically, the O_3 sensor outputs voltages named o3op1, o3op2 whereas the NO_2 sensor outputs voltages named no2op1, and no2op2.

The manufacturer of these sensors claims that these two voltages can give the true level of the pollutant using a simple linear model. However, these sensors are cross-sensitive in that the ozone sensor measures levels of not just ozone but also nitrogen dioxide. Thus, the manufacturer suggests that we use all 4 voltage values o3op1, o3op2, no2op1, and no2op2 along with a linear model to obtain the true value of both pollutants. Specifically, we wish to learn some real-valued constants p_{o3}, q_{o3}, r_{o3}, s_{o3}, t_{o3} such that the true level of ozone is given by

p_{o3}.o3op1 + q_{o3}.o3op2 + r_{o3}.no2op1 + s_{o3}.no2op2 + t_{o3}

and for some other real-valued constants p_{no2}, q_{no2}, r_{no2}, s_{no2}, t_{no2}, we have the true level of nitrogen dioxide given by

p_{no2}.o3op1 + q_{no2}.o3op2 + r_{no2}.no2op1 + s_{no2}.no2op2 + t_{no2}

2 Question-1

Find out how well can you predict the $O_3$and NO_2 using the method suggested by the manufacturer. To do this, learn the best linear model that uses just the 4 voltage values to predict O_3 and NO_2 values. Remember that for this part, you cannot use non-linear models, nor can you use temp, humidity, and timestamp 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.**

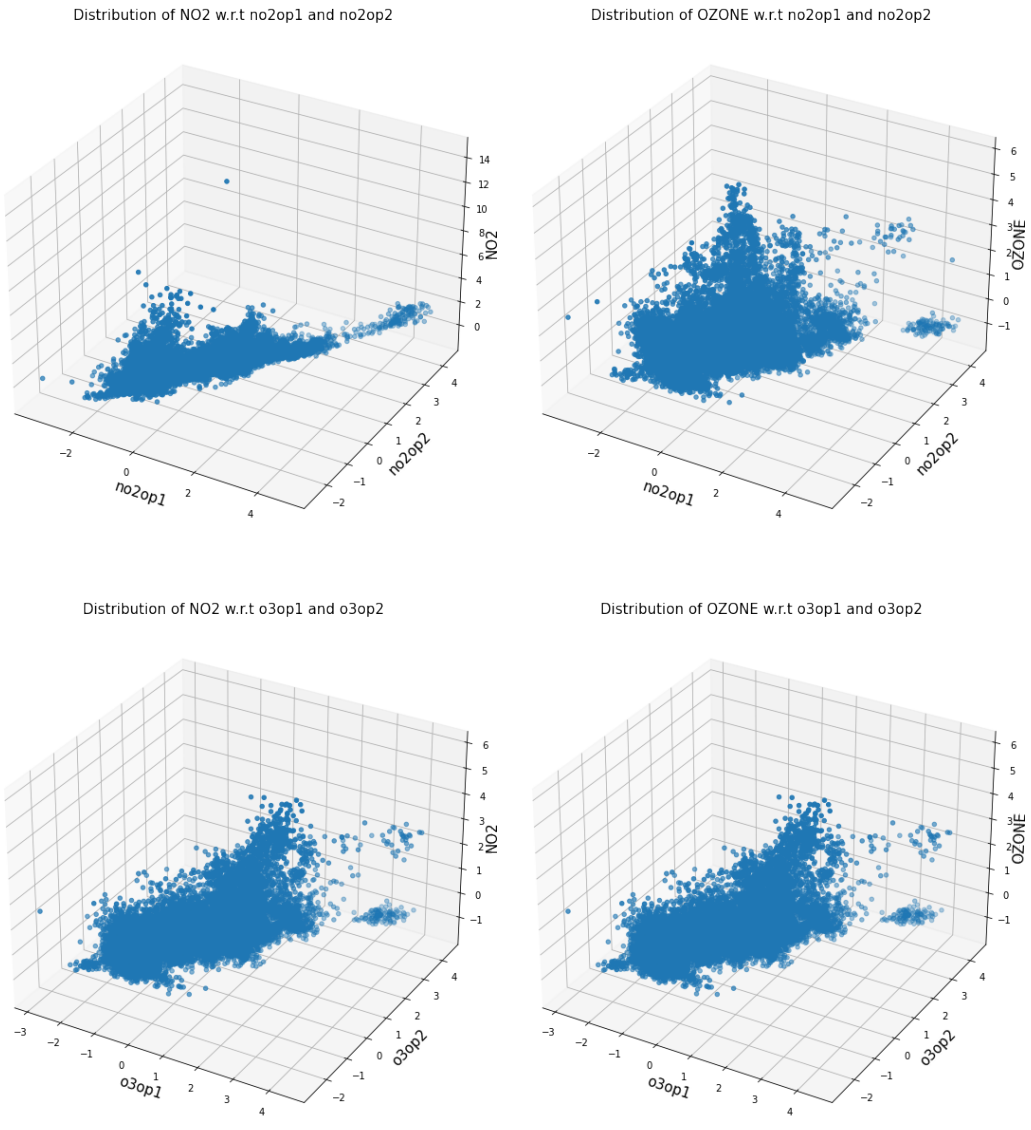
Solution:
The linear model we are using here presumes a linear relationship between input features and output variables. This indicates that the effect of changing a particular input feature on the output variable is proportional to the input feature's value and that a straight line or hyperplane can describe the overall relationship between the input features and the output variable.

However, the relationship between the input features and the output variable is non-linear; a linear model is incapable of accurately representing this relationship. But if the relationship between the input features and the output variable is curved or has a more complex form than a straight line or hyperplane, then the linear model we are using does not adequately represent the data.

This is due to the fact that linear models have limited flexibility and cannot capture the complex interactions between input features that may exist in a nonlinear relationship. Therefore, to model this nonlinear data, we require more sophisticated models, such as polynomial regression, decision trees, neural networks, or kernel methods, which can capture more complex relationships between input features and output variables.

The linear models are under fitted to the training data if it has low variance and high bias. Low variance indicates that the model's predictions are not sensitive to minor variations in the input data.

Underfitting occurs when the model is too simple to adequately represent the complexity of the data or when it has not been trained for enough epochs. Consequently, the model will not adequately match the training data and will also perform poorly on new, unseen data.



On the training data, we used SVR and linear regression. Some implications are as follows.

Model	MAE - OZONE	MAE - NO_2
Linear Regression	6.47354775896776	5.7544842786487065
Support Vector Regression(SVR)	6.294971056330985	5.843476672614188

We can infer from the results above that SVR with a linear kernel performs marginally better than Linear Regression in terms of "OZONE" prediction. While it is exactly the opposite for the prediction of NO_2. It might be because, for OZONE, SVR with linear kernel captures the relationship just a little bit better than linear regression, and, for NO_2 prediction, linear regression is just a little bit better than SVR with linear kernel.

3 Question-2

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, and timestamp in addition to the voltage features to predict the O_3 and NO_2 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.**

Solution:

- The fact that we have accepted that the training data is non-linear and cannot be trained using linear methods makes this question different from the prior one. Therefore, we used RandomForestRegressor with criterion = 'absolute error' and Support Vector Regression (SVR) with a non-linear kernel (RBF), as well as additional features like humidity, temperature, and time.
- Additionally, we have taken a reference point for the time feature, which is "2019-03-27 00:00:00," in order to measure a net time in seconds for the feature vector to use.
- As standardization can speed up the convergence of optimization algorithms used in training machine learning models, we also applied it to the feature vectors.
- First, we trained Support Vector Regression (SVR) over the modified data using a non-linear kernel ('rbf'). The hyperparameter (C) was tweaked, and the "minimum absolute error" was obtained at C=10.
- Results obtained on the validation script are predictionTime: 8.921952972999998, MAE-O_3: 3.9103548851479326, and MAE-NO_2: 3.0948332626190482.
- Second, we trained RandomForestRegressor with n_estimators = 10 over the modified data, and the results obtained on the validation script are predictionTime: 0.10493387940002777, MAE-O_3: 1.4326729500404183, and MAE-NO_2: 0.8616288096637039.
- Therefore, we have determined that the Random Forest Regressor is the best non-linear model in terms of performance over our training dataset, with the number of estimators equal to 10 and the criterion equal to "absolute error."