# CS771 Assignment 3

TANEYA SONI, ABHISHEK KUMAR PATHAK, VEDIT KUMAWAT, SONAM, ANIKET KUMAR

TOTAL POINTS

## 54 / 60

QUESTION 1

### 1 Description of best linear model 10 / 10

✓ **+ 6 pts** *Description of linear method*

✓ **+ 4 pts** *MAE of best linear method for both gases*

**- 2 pts** Incomplete or vague description e.g. MAE for both gases not reported separately

**+ 0 pts** Completely wrong or else unanswered

QUESTION 2

### 2 Description of best non-linear model 8 / 10

✓ **+ 10 pts** *Description of the best non-linear method*

✓ **- 2 pts** *Incomplete description e.g. missing best values for hyperparameters or not trying out reasonable hyperparameters e.g. depth of tree*

**+ 0 pts** Completely wrong or else unanswered

QUESTION 3

### 3 Experiments 36 / 40

**+ 0 pts** No submission

**+ 36** *Point adjustment*

GROUP NO: 53

Grading scheme for code:

Inference time ti (in sec): ti < 0.5 (10 marks), 0.5 <= ti < 1 (9 marks), 1 <= ti < 2 (8 marks), ti >= 2 (7 marks)

Submission size ss (in MB): ss < 2 (5 marks), 2 <= ss < 4 (4 marks), 4 <= ss < 8 (3 marks), ss >= 8 (2 marks)

MAE O3 mo: mo < 3.50 (10 marks), 3.50 <= mo < 4.00 (9 marks), 4.00 <= mo < 4.50 (8 marks), 4.50 <= mo < 5.00 (7 marks), 5.00 <= mo < 10.00 (6 marks), mo >= 10.00 (5 marks)

MAE NO2 mn: mn < 4.00 (15 marks), 4.00 <= mn < 5.00 (14 marks), 5.00 <= mn < 6.00 (13 marks), 6.00 <= mn < 7.00 (12 marks), 7.00 <= mn < 8.00 (11 marks), mn >= 8.00 (10 marks)

ti = 0.547 sec : 9 marks

ss = 12.489 MB : 2 marks

mo = 3.115 : 10 marks

mn = 1.83 : 15 marks

TOTAL: 36 marks

# Assignment-3


**Sonam**
221110057
sonamk22@iitk.ac.in

**Aniket Kumar**
190134
aniketkr@iitk.ac.in

**Vedit Kumawat**
190957
vedit@iitk.ac.in

**Taneya Soni**
221110062
taneyas22@iitk.ac.in

**Abhishek Kumar Pathak**
22111002
akpathak@iitk.ac.in


## 1 Question

*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, $\epsilon$ -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.*

We tried various Linear model such as SVR, OLS, Ridge, Lasso and tried hyper tuning there hyperparameters in order to get best fit as possible on the training data. The result for all the above 4 models are listed below in the table.

| Models | MAE( Ozone) | MAE( NO2) |
|---|---|---|
| OLS | 5.62 | 6.54 |
| Lasso | 5.65 | 6.57 |
| Ridge | 5.63 | 6.54 |
| SVR | 5.60 | 6.10 |

From the table we can see that SVR (linear Kernel) out perfomed all other model in the regression task giving us best MAE for both the tasks. It gave 5.60 mae on Ozone prediction and 6.10 on $NO_2$ prediction.

## 2 Question

*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 as well as the time stamp 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.*

## 1 Description of best linear model 10 / 10

✓ **+ 6 pts** *Description of linear method*

✓ **+ 4 pts** *MAE of best linear method for both gases*

**- 2 pts** Incomplete or vague description e.g. MAE for both gases not reported separately

**+ 0 pts** Completely wrong or else unanswered

### 2.1 solution

As we are allowed to use all the features to predict NO2 and OZONE, in order to do we used the 4 voltages as well as temperature, humidity and time stamp. To use time stamp we created 4 features out of it, namely the month, day, hour and minute. These were created so that decision tree based model can better learn using time stamp and also it convert string based time stamp in to 4 numerical value that each provide some important information. As we have already many lienar algorihtms in preivous question this time we focused on non linear models. We tried KNN regressor, Neural Networks, Random forest and XGBoost. In order to find the model that has the best predictive power we split the data into train and test(20% for test set) and checked which model performed best on the test set.

| Models | MAE( Ozone) | MAE( NO2) |
|---|---|---|
| XGBoost | 3.41 | 1.91 |
| Random Forest | 3.61 | 2.54 |
| Neural Nets | 5.14 | 5.98 |
| KNN-Regression | 3.98 | 2.98 |

From the table it is clear that XGBoost performed much better compared to other models. XGBoost works by iteratively adding new decision trees to the ensemble while adjusting the weights of the training examples to minimize the error of the overall model. The gradient descent optimization technique is used to minimize the loss function, and a regularization term is added to control over-fitting.
Now we then moved on to tuning the hyper parameters in order to do we first split the data into three set train(70%), valid(10%) and test(20%). Then we trained on training set and tuned hyper parameters so that it minimises the loss on valid set, be used grid search approach on n-estimators, learning rate, lambda and alpha parameter.
After that we were able to decrease the mae on Ozone to 3.33 and mae on $NO_2$ to 1.84 on the test set. With these hyper parameters we trained the model on the whole data set(100%) and submitted it for evaluation.

## 2 Description of best non-linear model 8 / 10

✓ **+ 10 pts** *Description of the best non-linear method*

✓ **- 2 pts** *Incomplete description e.g. missing best values for hyperparameters or not trying out reasonable hyperparameters e.g. depth of tree*

**+ 0 pts** Completely wrong or else unanswered

## *3* Experiments **36 / 40**

**+ 0 pts** No submission

**+ 36** *Point adjustment*

GROUP NO: 53

Grading scheme for code:

Inference time ti (in sec): ti < 0.5 (10 marks), 0.5 <= ti < 1 (9 marks), 1 <= ti < 2 (8 marks), ti >= 2 (7 marks)

Submission size ss (in MB): ss < 2 (5 marks), 2 <= ss < 4 (4 marks), 4 <= ss < 8 (3 marks), ss >= 8 (2 marks)

MAE O3 mo: mo < 3.50 (10 marks), 3.50 <= mo < 4.00 (9 marks), 4.00 <= mo < 4.50 (8 marks), 4.50 <= mo < 5.00 (7 marks), 5.00 <= mo < 10.00 (6 marks), mo >= 10.00 (5 marks)

MAE NO2 mn: mn < 4.00 (15 marks), 4.00 <= mn < 5.00 (14 marks), 5.00 <= mn < 6.00 (13 marks), 6.00 <= mn < 7.00 (12 marks), 7.00 <= mn < 8.00 (11 marks), mn >= 8.00 (10 marks)

ti = 0.547 sec : 9 marks

ss = 12.489 MB : 2 marks

mo = 3.115  : 10 marks

mn = 1.83  : 15 marks

TOTAL:  36 marks