Problem 3.1 (Calibrate Away). 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 O_3 and nitrogen dioxide NO_2 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, 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 manufacture suggests that we use all 4 voltage values o3op1, o3op2, no2op1, 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

$$p_{o3} \cdot \text{o3op1} + q_{o3} \cdot \text{o3op2} + r_{o3} \cdot \text{no2op1} + s_{o3} \cdot \text{no2op2} + t_{o3}$$

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

$$p_{no2} \cdot \text{o3op1} + q_{no2} \cdot \text{o3op2} + r_{no2} \cdot \text{no2op1} + s_{no2} \cdot \text{no2op2} + t_{no2}$$

Your Data. We have provided you with training data in a CSV file that contains 9 columns:

1. Timestamp: this tells us the time of the day at which the measurement was taken. One measurement was taken per minute
2. OZONE: this tells us the true level of O_3 at that time stamp
3. NO2: this tells us the true level of NO_2 at that time stamp
4. temp: this tells us the temperature at that time stamp in degrees Celcius (between 0 and 100)
5. humidity: this tells us the relative humidity at that time stamp as a percentage (between 0% and 100%)
6. no2op1, no2op2, o3op1, o3op2: these tell us the voltages given by the two sensors at that time stamp

Your Task. There are three tasks you have to perform

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, 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. (10 marks)

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 as well as the time stamp 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.** 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. (10 marks)

3. Use the training data to train your model on the expanded set of features and send us that model. Also write code that can take test features (timestamp, temp, humidity, no2op1, no2op2, o3op1, o3op2) and predict the value of O_3 and NO_2 using the model you have sent us. You are allowed to use all standard Python libraries e.g. numpy, sklearn, scipy, keras etc. **However, if you are using a non-standard library that is not available via pip and which is essential for prediction on test data, you must supply that library to us in your submission.** If a library is only needed for training and not for testing, then no need to submit that library. **Do not submit training code.** Submit prediction code for your chosen method in `submit.py`. Your code must implement a `my_predict()` method that takes a dataframe as input containing the test features and returns two numpy arrays, the first numpy array containing the predictions of O_3 for each test point and the other one containing NO_2 predictions on each test point. We will evaluate your submitted model on test data that is similar to the training data provided to you (see below for details). **Please go over the Google Colab validation code and the dummy submission file `dummy_submit.py`** to clarify any doubts about data formats, protocol etc. (40 marks)

Parts 1 and 2 need to be answered in the PDF file whereas Part 3 needs to be submitted as code + model. Remember, that part 1 can use only voltage values and linear models whereas in parts 2 and 3, you can use all features and non-linear models. Please submit only one model, the one that you feel is best. **Do not submit the linear model you found to work best in Part 1** (unless you could not find a non-linear model working better).

Evaluation Measures and Marking Scheme. We have secret test data that is similar to the training data and collected using the same two sensors that generated training data. We will evaluate your method on 3 criterion.

1. How fast is your `my_predict` method able to finish prediction (10 marks)

2. What is the on-disk size of your submission (after unzipping) (5 marks)

3. What MAE does your model offer for O_3 and NO_2 predictions on test data (10 + 15 marks)

Validation on Google Colab. Before making a submission, you must validate your submission on Google Colab using the script linked below.