Projects

Jose M. Barcelo Ordinas

Universidad Politècnica de Catalunya (UPC-BarcelonaTECH),
Computer Architecture Dept.
joseb@ac.upc.edu

May 5, 2022

1 Project 3. Calibration of sensors in uncontrolled environments in Air Pollution Sensor Monitoring Networks.

The objective of this project is to calibrate an air pollution sensor in an air pollution monitoring sensor network. We will take the data sampled of one IoT (Internet of Things) node that accommodates an array of sensors. The data captured is O3 (tropospheric ozone), NO2 (nitrogen dioxide), SO2 (Sulfur dioxide), NO (nitrogen monoxide), PM10 (Particulate Matter of size 10 micrometers), and two environmental parameters such as ambient temperature and relative humidity. The combined data set contains around one thousand samples.

You have to write a report with the main result of the calibration process for the O3 (tropospheric ozone). For that reason, the dataset provides the true value of O3 from a reference station. Write your findings, plot curves, and discuss the results. Here there is a description of some steps that you may follow.

2 Project Realization (Part I): observe the data (1 point)

The data consists on CSV files:

File "captor17013-sensor1.csv":

date;RefSt;Sensor_O3;Temp;RelHum
21/06/2017 7:00;15.0;36.3637;21.77;53.97
21/06/2017 7:30;15.0;34.8593;25.5;42.43

where it can be seen that the first row is a header that describes the data:

- **date:** Timestamp (UTC) for each measurement,
- **RefSt:** Reference Station O_3 concentrations, in μgr/m^3,
- **Sensor_O3:** MOX sensor measurements, in KΩ,
- **Temp:** Temperature sensor, in °C,

- **RelHum:** Relativa humidity sensor, in %.

Other files such as NO2_Manlleu.csv, NO_Manlleu.csv, SO2_Manlleu.csv and PM10_Manlleu.csv contains data at the same site with values for the other pollutants. I also provide you with a file called MLR_Build_File_Pandas-HW3.py to upload the data and build a PANDAS data frame where all data is merged. PANDAS is a python library for data manipulation and analysis that will help you during the project. The first step consists on understanding the data. For that purpose, the best approach is to plot several curves to see dependencies of the data.

1. The O3 sensor is a low-cost sensor (15 Euro) and works as a voltage divisor. That means, as explained in class, that it is represented as a variable resistor. Thus the first step is to plot the ozone (KOhms) as function of time to observe the data. Moreover, it also is interesting to plot the O3 reference data as a function of time and compare the sensor raw data with the reference data to see that they follow similar patterns.

2. In order to observe the linear dependence between the O3 reference data and the O3 low-cost sensor data, draw a scatter-plot, a plot in which in the x-axes you have the O3 low-cost sensor data and in the y-axes you have the O3 reference data. Ideally, the data should looks like linear with a tangent of 45 degrees. You will see that it is not a perfect line since the sensor is not perfect (thus the calibration). Sometimes the scatter-plot is difficult to observe due to the scale (ozone sensor is in Kohms and ozone concentration is in μgr/m^3), then you can normalize each data point with respect its mean and standard deviation. It is to say, for example to normalize the ozone sensor data: (i) Obtain the mean of the training set, μ_{sensor}, (ii) Obtain the standard deviation (std) of the training set, σ_{sensor}, and (iii) Normalize all the samples of the training: for j=1,..., K_1,

$$\bar{x}_{sensor_j} = \frac{x_{sensor_j} - \mu_{sensor}}{\sigma_{sensor}} \tag{1}$$

   where, $\bar{x}_{sensor_j}$ are the normalized sensor data. Do the same for the reference data and plot again the scatter-plot. In our case, you should not have difficulties in plotting the scatterplot in the original scale, thus, it is not necessary to normalize. In any case, you can do it in order to practice and see that the normalization does not change the pattern.

3. It is also interesting to plot scatterplots between the O3 low-cost sensor data and the rest of parameters (temperature, relative humidity, NO2, NO, SO2). Do the same with the O3 reference station.

At this moment you know the dependencies and can get some insight on the dependencies. You can also obtain useful information if you get the means of each parameter and a covariance matrix between all parameters.

3 Project Realization (Theoretical Part II): calibration.

Here you have a set of calibration methods to test and play. Make plots to explain results. You can obtain R^2, RMSE and MAE to check performance (see definitions in wikipedia), and also there are libraries in SciPy to get them. For each method you can plot the reference concentrations (e.g. in red) and the estimated concentrations (e.g. in blue), and you can regress (linearly) the Refdata against the estimation to see how good is the estimated data. At the end of the report, you can compare all the performance parameters (e.g. using a table) with all methods to see which method performs best.

1. Multiple linear regression (MLR): you have to perform a forward subset selection in order to identify the best features to take into account. Do not perform regularization at this step (the forward subset selection acts as regularization). Draw a table in which for each new feature you show the R^2, RMSE and MAE. Show also the optimal coefficients of the MLR. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

2. Multiple linear regression (MLR) with regularization: use both ridge regression and LASSO. Draw a table in which you show the R^2, RMSE and MAE and the regularization parameters. Show also the optimal coefficients of the MLR for each methods. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

3. K-nearest neighbor (KNN): for the best selection obtained in points 1, 2 or 3, select a set of features for performing a KNN regression. Show the R^2, RMSE and MAE. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

4. Kernel ridge regression with RBF kernel: for the best selection obtained in points 1, 2 or 3, select a set of features for performing a Kernel ridge regression. Show the R^2, RMSE and MAE. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

5. Random forest (RF): for the best selection obtained in points 1, 2 or 3, select a set of features for performing a RF regression. Show the R^2, RMSE and MAE. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

6. Support Vector Regression (SVR): for the best selection obtained in points 1, 2 or 3, select a set of features for performing a SVR regression with RBF Kernel. Show the R^2, RMSE and MAE. Plot the estimated O3 values against time and in the same plot add the O3 reference data.

7. Draw a table comparing all methods and discuss which method you prefer or performs better.