Homework 3: calibration of sensors

Jose M. Barcelo Ordinas, Jorge Garcia Vidal

November 24, 2023

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

The objective of this project is to calibrate an air pollution sensor. We will take the data sampled of one node that accommodates three sensors: a MIC2614 O_3 (ozone) sensor, a temperature sensor and a relative humidity sensor. The dataset contains a thousand samples. You have to write a report with the main results of the calibration process.

The theory related with the project are sections 7 to 12 of the LNs on Estimation, and also the sections on Linear Regression of the LNs on supervised ML.

2 Part I: Explore the data.

The data consists on a CSV file called "datos-17001.csv", being the format the following:

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.

The first row of the file 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:** Relative humidity sensor, in %.

The first step consists in uexploring the data. For that purpose, the best approach is to plot several curves to see dependencies of the data. We recommend using PANDAS as tool to handle the data.

1. The ozone sensor works as a voltage divisor. That means that it is represented as a variable resistor. Thus the first step is to plot the reading of the ozone sensor ($k\Omega$) as function of time to observe the data. Moreover, it is interesting to plot the ozone reference data as a function of time. You can compare them to check that they follow similar patterns.

2. In order to observe the linear dependence between the reference data and the sensor data, plot a scatter-plot, the x-axes corresponds to the ozone sensor data and the y-axes to the reference data.

   The ozone sensor readings are in $k\Omega$, while the reference stations reports ozone concentration measured in μgr/m^3. To compare both signals it is better to normalize them with respect to its mean and standard deviation:
   (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,

   $$\bar{x}_{sensor_j} = \frac{x_{sensor_j} - \mu_{sensor}}{\sqrt{\sigma^2_{sensor}}} \tag{2.0.1}$$

   where, $\bar{x}_{sensor_j}$ are the normalized sensor data.

   Note that with this operation, the normalized signals have a zero mean and standard deviation of 1.

   Ideally, the data points should follow a linear function with a slope of 45 degrees. In the real dataset, and due to measurement errors, the points appear in a cloud.

3. It is also interesting to plot scatterplots of the sensor with respect to temperature and with respect to relative humidity, and scatter plots of the reference station with respect to the temperature and with respect to the relative humidity. Normalize again all the data to have zero mean and standard deviation of 1.

3 Part II: Sensor calibration using multiple linear regression.

Yoy have to calibrate the sensor using a multiple linear regression model with three features. That means that:

$$\bar{y}_{RefSt_j} = \theta_0 + \theta_1 x_{s_{O3_j}} + \theta_1 x_{s_{Temp_j}} + \theta_1 x_{s_{RH_j}} + \sigma_j^2 \tag{3.0.1}$$

Use the library in python sklearn.

In order to obtain a Training set and a Validation set, *first randomly shuffle the data* and dedicate a percentage, the first 70% of the data set, for Training Set and

the rest (30%) for Validation Set. Other possibility is to use cross-validation. In any case, this data set is simple and it should work with a training/testing split.

You can use as metrics of your training/testing data set the coefficient of determination (R^2), and the Root Mean Square Error (RMSE), i.e. the square root of the MSE. Put a table with the estimated coefficients, and the value of the obtained metrics for training data and testing data.

You should also draw a plot with 2 curves: estimated sensor data as a function of time, and reference data as a function of time (each one with one color). Finally, draw a scatter plot of estimated sensor data against reference data (and add a line 45^{o}). Comment your results.