Projects

Statistical Analysis of Networks and Systems (SANS-MIRI)

December 21, 2021

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 in an air pollution monitoring sensor network. We will take the data sampled of one node that accommodates three sensors: a MIC2614 O_3 sensor, a temperature sensor and a relative humidity sensor. The data set contains one thousand samples. You have to write a report with the main results of the calibration process.

2 Project Realization (Part I): observe 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

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:** Relative humidity sensor, in %.

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. I 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 ozone (KOhms) 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 and see that they follow similar patterns.
2. In order to observe the linear dependence between the reference data and the sensor data, draw a scatter-plot, a plot in which in the x-axes you have the ozone sensor data and in the y-axes you have the 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 of the sensor with respect to temperature and with respect to relative humidity. Do the same with the reference station with respect to the temperature and with respect to the relative humidity.

3 Project Realization (Part II): calibration using multiple linear regression (frequentist framework).

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

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

You can use the library in python sklearn. It is important that you shuffle the data and dedicate a percentage (e.g. 70

You can use as metrics of your training/testing data set the coefficient of determination (R^2), the root mean square error (SMRE) and the mean absolute error (MAE). 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 45Â°). Comment your results.

4 Project Realization (Part III): calibration using multiple linear regression (Bayesian framework).

In this case you repeat the calibration assuming, as seen in class, that the output follows a normal distribution of mean $\theta_0 + \theta_1 x_{s_{O3}} + \theta_1 x_{s_{Temp}} + \theta_1 x_{s_{RH}}$ and variance σ_j^2. Now you have to obtain the posterior of the parameters θ using a Markov Chain Monte Carlo (MCMC) simulation. For that use, as explain in class, the Generalized Linear Model (GLM) module from PyMC3.

As results, you should plot the histogram of the posterior parameters (θ_0, θ_1, θ_2, θ_3, and σ^2) and obtain their statistics (mean and variance).

Finally, use the mean of the posterior to estimate in the testing the ozone concentrations and obtain the R^2, RMSE and MAE, and plot the results against time and scatter plot as you did in the frequestist case.

Explain your final conclusions with respect to the two models.