SANS2023-24-HW4

José María Barceló, Jorge García Vidal

December 2023

In this HW we will use the same dataset as in HW3. Perform the normalization/denormalization of data as described in HW3.

1 Calibration using a FNN

LNs on ML, published on Nov 20:

- Sections 5 (FFNN),
- 7 (training ML),
- 8 (Approx and estimation errors) ,
- 9 (Regularization),
- 11 (underfitting overfitting),
- 12 (MSE),
- 14 (why regularization?),
- 15 (Model comparison),
- 16 (Training, Validation and testing sets), and
- 17 (CV)

LNs on ML published on Dec 28:

- 7(FFNN)
- 10 (training FFNN),
- 11 (Approx and estimation)
- 12 (Regularization)
- 14 (undefitting/overfitting)
- 15 (MSE Bias variance)

- 17 (Why regularization?)
- 18 (Model selection)
- 19 (Training Testing Validation sets)
- 20 (CV)

Calibrate the same sensor that in homework 3, but now using a FFNN. Use the same data, and use the pytorch library.

The input of the FFNN is the features (raw ozone measurements, temperature, and relative humidity). Use an architecture with an input layer, a hidden layer, and an output layer. Try using 2, 3, 4, and 10 neurons for the input and hidden layer. For the input and hidden layer use ReLU activation layer. Use a linear activation for the output layer.

Shuffle the data before calibration. Use 80% of data for training, 10% for validation, and 10% for testing. Plot the RMSE (training, and validation) for the four proposed architectures (2, 3, 4, and 10 neurons for the input and hidden layer).

2 Calibration using Bayesian multiple linear regression (LNs Bayesian estimation and ML, sections 1 to 10)

Perform the calibration of the sensor using Bayesian Linear regression assuming that that the output follows a normal distribution of mean $\theta_0 + \theta_1 x_{s_{O3}} + \theta_2 x_{s_{Temp}} + \theta_3 x_{s_{RH}}$ and variance σ^2. Now you have to obtain the posterior of the parameters θ and σ^2 using a Markov Chain Monte Carlo (MCMC) simulation. For that use, the Generalized Linear Model (GLM) module from PyMC3.

For the dataset, use the rows of data of file *datos-17001.csv* between June 21th and July 17th, i.e. from the second row:

$$21/06/2017 \;\; 7:00; 15.0; 36.3637; 21.77; 53.97$$

to row 943:

$$10/07/2017 \;\; 23:30; 73.0; 254.599; 22.0; 36.43$$

Plot the histogram of the posterior parameters (θ_0, θ_1, θ_2, θ_3, and σ^2) and report obtain their mean and variance.

Assume that the sensor has the readings that correspond to time 11/07/2017 11:00:

$$264.744; 28.97; 30.33$$

Plot the histogram of the predictive distribution. (The ref station reads 82.0)

Repeat for the readings at 11/07/2017 18:00:

$$498.9; 33.2; 28.3$$

(the ref station reads 164.0)

Plot the expected value of the predictive distribution and the credibility interval ($+/-2\sigma$) for the day July 11th (i.e. from row 944 to the end of the file), together with the values measured by the reference station.