Smart emission paper – Data calibration part

Raw Jose measurements of gas components are unreliable and interpreting these measurements is hard. At least three distortive aspects of the gas measurement need to be taken into account. First of all, the gas measurements are noisy, i.e. consecutive measurements in a short period of time are different. Secondly, the measurement are biased because the gas sensors are cross sensitive for meteo conditions and other gas components. Thirdly, the gas measurements are measured in kOhm instead of ug/m3 or ppm. One way to solve these three problems is calibration with ground truth data.

Ground truth data is recorded by the “landelijk meetnetwerk luchtkwaliteit” (lml) of the RIVM[[1]](#footnote-2). This network of air quality stations is used to report about the air quality in the Netherlands. Two Jose sensors are mounted on top of two RIVM air quality stations that are located in Nijmegen. Data was gathered from February until August and in this period the temperature ranged between 0C and 30C.

A feed-forward neural-network [1] is used to find the relation between the Jose and RIVM measurements in the calibration data. A feed-forward neural-network can learn non-linear relations between the input and output data that cannot be found with linear regression. Specifically, the MLPRegressor[[2]](#footnote-3) implementation from the scikit-learn package is used. The calibration data is made up out of seven normalized Jose measurements (humidty, pressure, internal and external temperature, CO2 in ppm, O3 in kOhm and NO2 in kOhm). The target data is one of the lml measurements of CO, O3 or NO2.

The neural-network can be trained if the network architecture is given. However, depending on the size and number of hidden layers and other parameters the performance might vary. For this reason, different network architectures and other model learning parameters are compared. A random cross-validated search is done to find good performing sets of parameters. The set of parameters with the most explained variance (r^2) is chosen. Ranges for the parameters can be found in Table 1. The final parameters that are used to train the model are in Table 2.

The cross-validation results suggest that O3 can be reliably predicted (r^2 = 0.91, SD = 0.02; RMSE = 7.48ug/m3, SD = 1.06 ug/m3) and that the results of NO2 (r^2 = 0.51, SD = .03; RMSE = 12.57 ug/m3; SD = 0.67 ug/m3) and CO (r^2 = 0.49, SD = 0.06; RMSE = 124.45 ug/m3, SD = 8.91 ug/m3) are indicative for ground truth values. Scatter plots of the predicted and actual values of the gas components are shown in Figure 1-3.

**References**

[1] Hinton, G. E. (1989). Connectionist learning procedures. Artificial intelligence, 40(1), 185-234

**Tabel 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Component | Parameter | Description | values |
| mlp | hidden\_layer\_sizes | The size of the hidden layer. With a larger hidden layer size the network can learn more complex functions. | U(2, 150) |
| mlp | activation | The non-linear activation functions used for the hidden layer. Logistic and tanh squash the activation between a discrete range. Relu is linear above zero and constant below zero. | {logistic, tanh, relu} |
| mlp | alpha | Regularization term for the weights. A higher value forces weight more towards zero and is useful for preventing over-fitting. | 10^U(-6, -1) |
| mlp | learning\_rate\_init | The size of the change after each learning iteration. A higher learning rate changes the network faster but might never converge to good solutions. | 10^U(-6, -1) |
| filter | alpha | The weight of a new observation in the running mean of the gas components. A lower value is more robust against outliers but less sensitive to temporary changes in the environment. | 10^U(6, 0) |

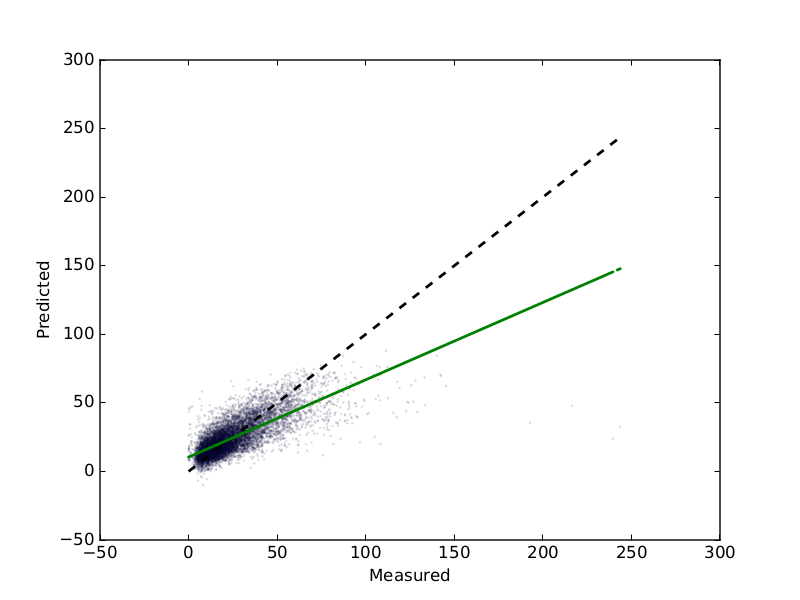
Different parameters that were randomly sampled and tested. U(a, b) is a uniform distribution between a and b, {x, y, ...} is a choice from x, y or something else, 10^U(a, b) is a exponential distribution between 10^-6 and 10^-1. The

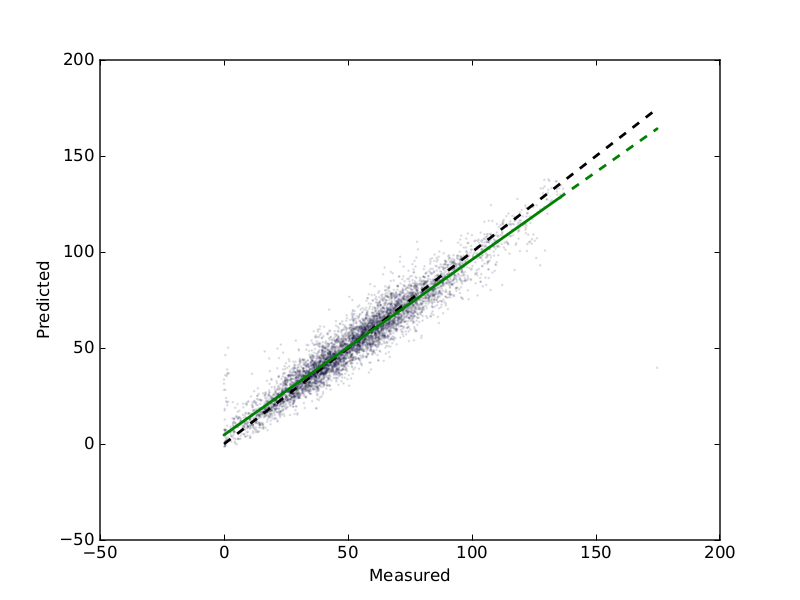
**Tabel 2**

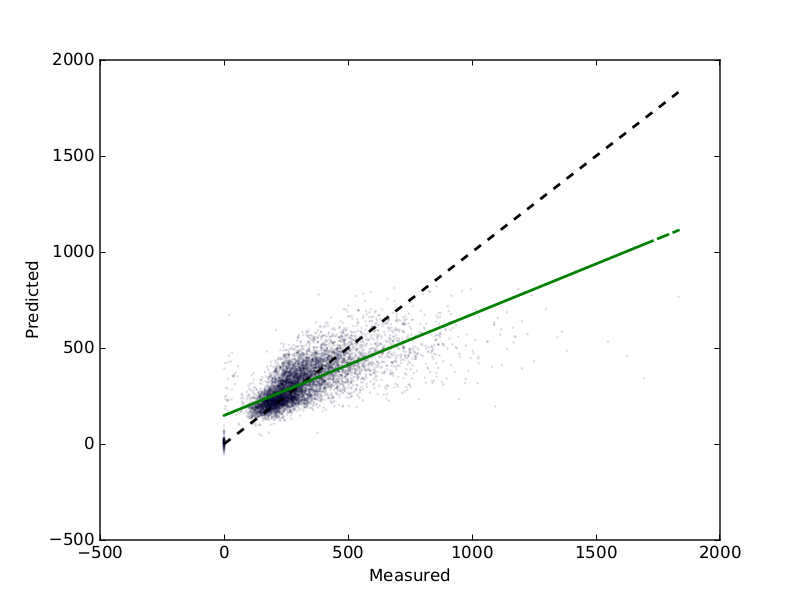
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Component | Parameters | O3 | CO | NO2 |
| Mlp | hidden\_layer\_sizes | 141 | 112 | 71 |
| Mlp | Activation | logistic | relu | tanh |
| Mlp | Alpha | 0.000014 | 0.0000018 | 0.000014 |
| Mlp | learning\_rate\_init | 0.0045 | 0.079 | 0.000042 |
| Filter | Alpha | 0.0077 | 0.007 | 0.013 |

Final parameters chosen for calibration of gas components O3, CO and NO2.

**Figure 1: NO2**



**Figure 2: O3**

**Figure 3: CO**

1. See <http://www.lml.rivm.nl/meetnet/index.php> [↑](#footnote-ref-2)
2. See [http://scikit-learn.org/dev/modules/generated/sklearn.neural\_network.MLPRegressor.html#sklearn.neural\_network.MLPRegressor](http://scikit-learn.org/dev/modules/generated/sklearn.neural_network.MLPRegressor.html" \l "sklearn.neural_network.MLPRegressor) [↑](#footnote-ref-3)