TASK 2: Regression

Pre-processing

All ten batches will be combined into a single data set. An additional column 'BATCH' denoting the batch to which the observation belongs will be added. Next, all features will obtain an appropriate column name.

Data exploratory analysis

First, the general characteristics of the dataset itself will be investigated. The dataset its shape and the number of observations for each gas will be requested. Next, the distribution of the target variable will be analysed. This by plotting the targets variable global distribution and the distribution of the target variable for each particular gas. Scatterplots representing the mean and standard deviation of the features of the full dataset and for each particular gas will be created. Next, the correlation between each feature pair will be computed. At last, an isolation forest model will be fitted to obtain an estimate of the outlier proportion. The observations that are identified as outliers will be further investigated. To be precise, it will be examined whether there is a particular gas or particular concentration that frequently is identified as an outlier.

Research

There are three tasks to fulfil:

1. Train a separate regression model for every gas
2. Train a single regression model to estimate the concentration of any gas
3. Train a single regression model to estimate the concentration of all 6 gases separately

5 different models will be learned for each task. The performance of each model will be expressed by the use of MSE, MAE and R2 metric. These models will first be trained on data without feature selection. Next, the same machine learning models will be trained on the same data but after the application of different dimensionality reduction mechanisms. The same performance metrics will be computed to compare performance with the baseline models.

The research starts with the creation of 8 different datasets, 1 dataset without the gas feature, 1 dataset that includes a categorical gas feature and 6 datasets only containing observations of one particular gas. Each dataset will be divided into a (0.75-0.25) training/test set. An Elastic Net, k-Neighbors, SGD, AdaBoost and Bayesian Ridge regression model will be trained on each training set. Model hyperparameters will be tuned by performing a 5-fold-cross-validation grid search on potential values. In case the amount of hyperparameters combinations is too large to evaluate separately, a random search will be used instead. The created models will predict the target variable for each test set observation. Finally, the MSE, MAE and R^2 will be computed for each created model.

This procedure will be repeated but now after the application of a dimensionality reduction algorithm. A principal component analysis, recursive feature selection and the feature importance attribute of a random forest will be used for this task. These algorithms will only be learned on the training set but applied on both training and test set.

Each model trained for task two will contain a dummy coded categorical gas feature. In case this feature is removed due to regularization or dimension reduction, the feature will be added again before model training to assure that the model is able to fulfil the requirements of task two.

Note: If an algorithm uses a distance metric the data will first be normalized.