Regression Assignment: Predicting pollution levels

1 Task Description

The health department of a region has noticed that pollution indicators have drastically increased. The regional government, concerned about this situation, has decided to implement a prediction model to predict the concentration on carbon monoxide in the air from the date gathered by the air-quality monitors.

In this assignment, you will use regression techniques in order to predict the CO concentration based on temporal, climatic and environmental data.

The quality of your prediction is evaluated according to the Mean Absolute Error (MAE) metric.

2 Dataset Description

The dataset contains hourly information of weather and environmental conditions in a time period from 2014 to 2018.

The target attribute is the carbon monoxide concentration (`carb_monox`)

Number of instances: 14000

Number of attributes: 12

2.1 Target Class:

`carb_monox`: CO concentration (micrograms per cubic meter)

2.2 Attribute Information:

Item	Attribute	Type	Values
1	**hr:** hour of the day	Numerical/Categorical	0-24
2	**small_part:** fine particulate matter concentration	Numerical	micrograms per cubic meter
3	**med_part:** particulate matter concentration	Numerical	micrograms per cubic meter
4	**sulf_diox:** Sulfur dioxide concentration	Numerical	micrograms per cubic meter
5	**nitr_diox:** Nitrogen Dioxide concentration	Numerical	micrograms per cubic meter
6	**trioxygen:** Ozone concentration	Numerical	micrograms per cubic meter
7	**temp:** Temperature	Numerical	degree Celsius
8	**pres:** Pressure	Numerical	hectopascals
9	**rain:** Precipitation	Numerical	millimeters
10	**wind:** Wind direction	Categorical	E, ENE,ESE,N,NE,NNE, NNW,NW,S,SE,SSE,SSW, SW,W,WNW,WSW
11	**wind_sp:** Wind speed	Numerical	meters per second
12	**date:** Date	Categorical	day-month-year

3 Predictions

The prediction of your model can be evaluated with the following datasets:

- prediction_dependent: contains the explanatory variables of 6000 new records.
- prediction_independent: contains the real value of the previous observations that you can confront with your predictions.