plume labs

Spatiotemporal PM10 concentration prediction

Contact

Grégoire Jauvion, Lead Data Scientist – gregoire@plumelabs.com

Company description

Plume Labs is a technology company providing air quality live data and forecasts to urban consumers and businesses.

Our mobile app, the Plume Air Report, provides air quality levels around the world to consumers. Our personal air quality tracker Flow senses pollutants around you to help you avoid them at home and on the go – and crowdsource highly valuable hyperlocal maps in the process. Our global atmospheric pollution API gives businesses and academic teams an access to our unique AI-powered air quality forecasts data platform.

Plume Labs is an MIT/Stanford start-up with a team of 25 in New York and Paris and raised $4.5M in seed funding to date.

Problem definition

In order to provide air quality forecasts, Plume Labs has built a unique database with readings collected by monitoring stations all over the world. The problem we submit consists in predicting the PM10 concentration measured at some stations using urban features evaluated at the station location as well as the readings provided by the stations nearby.

The dataset gives the measures provided by 116 stations across Europe:

- 85 stations form the training dataset
- 31 stations form the test dataset

For each PM10 reading, the dataset provides the following information:

- Land-use characterization of the station location (i.e. is it located in a residential area, industrial area, …)
- Roads density evaluated at the station location
- PM10 readings at the closest stations as well as the distance to these stations (note: only the 85 stations of the training dataset have been used in this step, to ensure that the values to predict cannot be identified in these features)

The accuracy obtained by such a prediction model is a very good indicator of how an air quality prediction model performs in locations where there is no monitoring station.

Input and output variables description

Input variables:

- **Land-Use features at the monitoring station**
  - The land-use classification used is the following:
    - $hdres$: high-density residential area
    - $ldres$: low-density residential area
    - $industry$: industrial area
    - $urbgreen$: green area (mostly parks)
  - The number gives the radius of the area used to compute the feature (in meters)
  - Example: industry_500 gives the share of industrial areas at less than 500 meters to this station
  - All features are comprised between 0 and 1
- **Roads features at the monitoring station**
  - Two roads features are given:
    - $roads\_length$: length of all roads around
    - $major\_roads\_length$: length of major roads around
  - The classification roads / major roads is based on OpenStreetMap classification
  - The number gives the radius of the area used to compute the feature (in meters)
  - These features are normalized by the size of the area (hence, the unit is arbitrary)
  - Example: roads_length_500 is proportional to the length of roads at less than 500 meters to this station
- **Readings at the N closest monitoring stations**
  - $(value_0, \dots, value_9)$ give the readings at the 10 closest monitoring stations, sorted by decreasing distance
  - $(distance_0, \dots, distance_9)$ give the distances to the 10 closest monitoring stations, sorted by decreasing distance
  - Only the stations forming the training dataset are used in this step
  - Some values may be nan, which means that the corresponding monitoring station did not return any value at this moment

Output variable:

- The PM10 reading y

Benchmark

We propose a benchmark where the PM10 reading of a monitoring station is predicted using a weighted average of the readings provided by the stations nearby at the same time, with weights inversely proportional to the distance to the station.

$$y_{benchmark} = \sum_{i=0}^{9} \left(\frac{1}{distance_i}\right) value_i \,/\, \sum_{i=0}^{9} \left(\frac{1}{distance_i}\right)$$

Comparing the performance of a prediction model with the performance of this benchmark is extremely valuable, as it enables to quantify how using urban features improves the accuracy of the prediction.

Metric

The predictions are evaluated using the RMSLE metric:

$$Metric = \frac{1}{N}\sum_{i=1}^{N} (\ln(y_i + 1) - \ln(\overline{y_i} + 1))^2$$

, where y_i and $\overline{y_i}$ are respectively the true reading and the reading predicted by the participant.