Homework 3 y 4: black carbon proxy sensor in air quality sensor monitoring networks.

Jose M. Barcelo Ordinas

Universidad Politècnica de Catalunya (UPC-BarcelonaTECH),
Computer Architecture Dept.
jose.maria.barcelo@upc.edu

May 2, 2023

Black carbon is a major component of fine particulate matter, a potent warming agent in the atmosphere which contributes to regional environmental disruption and accelerates glacier melting. BC appears from incomplete combustion and comes mainly from road traffic, it is present in urban aerosols and is linked to cardiovascular and respiratory diseases. The monitoring of BC is not easy and is not regulated by the European Union (EU) Air Quality Directives. In contrast to regulated pollutants, there is not much affordable equipment to monitor BC, which makes the availability of BC measurements operationally expensive and difficult. The new proposal for a Directive of the EU Parliament on ambient air quality and cleaner air for Europe, published in October 2022, states that introducing additional sampling points for unregulated air pollutants of emerging concern, such as ultrafine particles, black carbon, ammonia or the oxidative potential of particulate matter (PM), will support scientific understanding of their effects on health and the environment. This supports the need to determine BC concentrations, either by direct or indirect methods, in European urban areas. In recent years there has been a great interest in using low-cost sensors (LCSs) to measure regulated pollutants such as CO, NO_2, NO, SO_2, O_3, and PM_{10}, $PM_{2.5}$ particles. One way to increase the availability of BC measurements without the need to deploy expensive equipment is the use of virtual sensors. A virtual sensor is defined as a mathematical model that estimates the target phenomenon at a specific location where no physical sensor is available. A proxy is a specific type of virtual sensor that estimates a target pollutant from indirect sensor measurements.

1 Homework 3: BC proxy using machine learning

The objective of homework 3 is to build a BC proxy using machine learning tools. The estimation model is non-linear, so, you can use any non-linear model. For this homework, we propose you to compare the proxy results with two well-known machine learning regression tools, support-vector regression (SVR) and random forest (RF). Here there is a description of some steps that you may follow.

The data consists on a CSV file: "BC-Data-Set.csv", where it can be seen that the first row is a header that describes the data:

- **date:** Timestamp (UTC) for each measurement,

- **BC:** true values of BC concentrations, in μgr/m^3,
- **N_CPC:** ultrafine particle number concentration,
- **PM-10:** sensor measurements, in μgr/m^3,
- **PM-2.5:** sensor measurements, in μgr/m^3,
- **PM-1:** sensor measurements, in μgr/m^3,
- **NO:** sensor measurements, in μgr/m^3,
- **O_3:** sensor measurements, in μgr/m^3,
- **SO_2:** sensor measurements, in μgr/m^3,
- **CO:** sensor measurements, in μgr/m^3,
- **NO:** sensor measurements, in μgr/m^3,
- **NOx:** sensor measurements, in μgr/m^3,
- **TEMP:** temperature sensor, in $^\circ$C,
- **HUM:** relative humidity sensor, in %.

The first step consists on understanding the data. For that purpose, the best approach is to obtain statistic (means, correlations, etc) and to plot several curves to see dependencies of the data (scatter-plots and temporal trends).

Now, the second step is to produce the proxy using SVR and RF and compare results in terms of R^2 and RMSE. However, you have too many input parameters and it is possible that some of the features are not useful at all. A good approach is to regularize the model. Compare the results using a forward subset selection mechanism or any regularization that fits you with both machine learning models. Again, plot results or/and use tables to show your results. Remember to optimize the hyperparameters of the models. You are also free to use other machine learning models if you wish, or to change kernels.

Plot results, show results, show intermediate values (e.g. results with several features using the subset selection), etc. Get your conclusions. Here, it is not so important the temporality of your data, so you can shuffle the data if you want to improve the results. Try both cases, shuffling and not shuffling. In either case, recover the temporality if you plot the results.

2 Homework 4: BC proxy using deep learning

The objective of homework 4 is to build a BC proxy using deep learning tools. In this case, build your proxy using an artificial neural network (ANN). Write your findings (describe the architecture you use justifying why you choose it), plot curves, and discuss the results with respect the results you obtained in homework 3.

As a second part, build a LSTM network to predict your results. Be careful with the temporality of the data. It is to say, break the training and testing in such a way that the temporality is maintained. If not, the LSTM is not going to do anything.

As a remark, I have not tested the LSTM on this dataset, so I do not know what to expect from the LSTM network. In any case, from a theoretical point of view, black carbon is highly dependent on particulate matter, a pollutant whose behavior is local and non smooth. In addition, the data set has gaps (missing values), so that certain parts of the data have lost temporality, as the corresponding rows have been deleted when a gap appeared. This tells us that it is rather debatable whether temporality and thus the LSTM will improve the prediction. However, it is a good exercise to see what happens. In any case, you can rerun the NN and LSTM being careful with the temporality of the data.