LELEC2870 Project

Prediction of air-quality in Beijing

Antoine Vanderschueren
antoine.vanderschueren@uclouvain.be - Stevin a.178

Niels Sayez
niels.sayez@uclouvain.be - Stevin a.156

Simon Carbonnelle
simon.carbonnelle@uclouvain.be - Stevin a.178

November 29, 2019

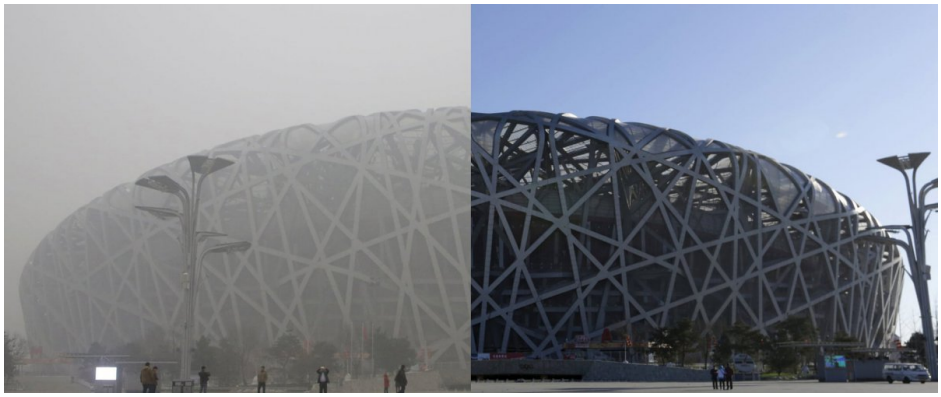

Figure 1: Air-quality is sometimes very problematic in Beijing [1].

Introduction

Machine learning methods can be used to solve many practical problems in a wide range of applications such as weather forecast, customer clustering, medical diagnostics, spam blocking, financial time series prediction or signal de-noising, ... In this project, you will apply machine learning to predict the air-quality in Beijing based on the concentration of certain particles, the temperature, precipitation, wind speed etc.

Data

The dataset you'll use contains hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017 [2].

You'll find the data on the course website in two csv files called X1.csv, Y1.csv and X2.csv. The first one is the labeled dataset and the second will be used to make your prediction. Paste these files in your working directory. The data can be loaded in your workspace by running the following commands:

```
import pandas as pd

# Use pandas to load into a DataFrame
#     Y1.csv doesn't have a header so
#     add one when loading the file
X1 = pd.read_csv("X1.csv")
Y1 = pd.read_csv("Y1.csv", header=None, names=['PM2.5'])

# If you prefer to work with numpy arrays:
X1 = X1.values
```

Each row in the data corresponds to an observation (a measurement). Each column represents one of the features listed here :

1. year: year of data Input Variable
2. month: month of data Input Variable

3. day: day of data Input Variable
4. hour: hour of data in this row Input Variable
5. SO2: SO2 concentration (ug/m^3) Input Variable
6. NO2: NO2 concentration (ug/m^3) Input Variable
7. CO: CO concentration (ug/m^3) Input Variable
8. O3: O3 concentration (ug/m^3) Input Variable
9. TEMP: temperature (degree Celsius)
10. PRES: pressure (hPa)
11. DEWP: dew point temperature (degree Celsius)
12. RAIN: precipitation (mm)
13. wd: wind direction
14. WSPM: wind speed (m/s)
15. station: id of the air-quality monitoring site
16. PM2.5: PM2.5 concentration (ug/m^3) Output Variable

Instructions

The project is realised by groups of two or alone. It is composed of different aspects as specified below.

Model

You will build regression models that predict the concentration of PM2.5 in the air of Beijing. You can use any of the methods seen during the lectures. We expect you to, at least, implement linear regression, KNN[1] and one other non-linear method. Features selection and model selection shall also be part of your work. Pay attention to the fact that the model selection can require a lot of computation time. You are advised to explore the metaparameters space according to the time available.

[1]K-Nearest Neighbour : Regression model with metaparameter K that predicts the output of a sample as the mean of output of the K nearest neighbours in the features space.

Prediction

Once your model is properly selected and validated, you are asked to produce predictions Y2 on the data X2 for which we have kept secret the corresponding targets. This prediction vector should be uploaded on Moodle in a csv file named "Y2.csv" that contains one line per prediction and no header. Your prediction quality criterion shall be the root mean squared (rms) error. In addition, you will provide in your report, an estimate of the rms error that you expect on your prediction.

Report

You will produce a report documenting your technical choices and experimental results. **We do not need a course on the methods you use. We are more interested in what you did and why.** Try to illustrate your results with graphics (with legends) and comment them. Be critical about what you observe and try to give a possible justification of the obtained results. Summarize your results and observations in a conclusion. A strict maximum of 7 pages (font of size 11 or larger) will be observed. Annexes might be included in the digital version that you'll submit on Moodle.

The report will be the basis of a discussion which will take place during the exam session. All your figures and computation need to be reproducible by us running your implementation code on the provided data.

Programming languages

The programming language you will use is Python. You can use any toolbox/library available on-line. In particular, we strongly recommend using the *scikit-learn* library as it provides many useful implementations of standard machine learning approaches.

Agenda

- As soon as possible: Register your group (maximum two people) on course website.
- Thursday December the 5th and Thursday December the 12th at 4:15pm: We will be available in the lecture room (BA91) to answer your questions.
- Friday 20th of December, 23h50 (to avoid any confusion on Moodle)
  - submit your work as an archive (.zip) containing the following items
    * Your report (pdf)
    * A csv file called "Y2.csv" no header line: one line per prediction values.
    * A folder containing all scripts you wrote for the project. The code should be commented well enough and installation instructions about non-standard packages you used should be provided.

Evaluation Criteria

- Respect of the instructions and deadlines
- Quality of the report and its defense (discussion)
- Your machine learning approach (model choices and validation)
- Reproducibility of your results
- Consistency between the report, your implementation and your predictions

Please note that the performance of your models is not critical in the evaluation.

Tips

Here is a list of advices for the project.

Before any analysis:

- Visualizing the data is always useful.
- Think about how to encode discrete or cyclic input variables: month, hour, wind direction... e.g. the 12th month is close to the 1st one, but when encoded simply with numbers from 1 to 12 this relation is not clear
- Normalize your data if necessary.

You can discuss the project with other students, in fact, it is a great idea! You could compare your results to those obtained by other groups, but remember that it is not allowed to copy what others did...

We will be happy to answer your questions during the Q/A sessions or on appointment. Good luck !

References

[1] Beijings air quality improvements are a model for other cities, CCAC secretariat 9 March, 2019
https://www.ccacoalition.org/en/news/beijing%E2%80%99s-air-quality-improvements-are-model-other-cities

[2] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S.X. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, Volume 473, No. 2205, Pages 20170457.