Task Description

Groups are requested to predict concentration levels of several pollutants over the coming 24*2 hours (two days) for 35 stations in Beijing, China.

We will provide air quality data and meteorological(weather) data from January 2017 to April 30 (including), 2018, and you need to predict the pollution level of PM2.5, PM10, O3 between May 1 to May 2, 2018 (once an hour, 48 times for one station in total).

Files Description

We provide the following data:

1.Air Quality Data (on hourly basis)

The Air Quality Data file contains the concentration of PM2.5 (ug/m3), PM10 (ug/m3), NO2 (ug/m3), CO (mg/m3), O3 (ug/m3) and SO2 (ug/m3) from Beijing. Your group only need to predict the concentration of PM2.5, PM10 and O3.

Related Files:

1) **airQuality_201701-201801.csv** From January 2017 to January 2018
2) **airQuality_201802-201803.csv** From February 2018 to March 2018
3) **airQuality_201804.csv** April 2018
4) **Beijing_AirQuality_Stations_en.xlsx** Location of air quality station in Beijing

2.Weather Data (on hourly basis)

The Weather Data contains 2 types of data: Observed Weather Data and Grid Weather Data.

We provide 651 points of grid weather data in Beijing and observed weather data from 18 weather stations in Beijing. Please check detailed information in **weatherData_detail.docx**

Related Files :

5) **observedWeather_201701-201801.csv** From January 2017 to January 2018
6) **observedWeather_201802-201803.csv** From February 2018 to March 2018
7) **observedWeather_201804.csv** April 2018
8) **observedWeather_20180501-20180502.csv** From May 1, 2018 to May 2, 2018(for prediction)
9) **gridWeather_201701-201803.csv** From January 2017 to March 2018

10) **gridWeather_201804.csv** April 2018

11) **gridWeather_20180501-20180502.csv** From May 1, 2018 to May 2, 2018(for prediction)

12) **Beijing_grid_weather_station.csv** Location of grid weather station in Beijing

13) **weatherData_detail.docx** Detailed information about weather data

3. Other Files:

14) **sample_submission.csv** Format of submission (the predicting results)

15) **evaluation.py** You can use this python file to calculate your score.

16) **station_map** This is a folder containing a html file. You can open it and check the detailed location of stations visually, including grid weather points, air quality stations and observed weather stations.

Please note that external data is allowed to use for this project.

Evaluation

The project is scored by report, code and predict results.

We will calculate your score of predict results via Symmetric mean absolute percentage error:

$$smape = \frac{1}{n}\sum_{t=1}^{n}\frac{|\mathrm{F}_t - \mathrm{A}_t|}{(\mathrm{A}_t + \mathrm{F}_t)/2}$$

Please note

1) for all the NA values in submission files, we will replace them by 0.

2) The submission file shall not contain any negative value.

3) The smaller score values indicate better performance and you can get score through **evaluation.py**

Submissions

You need to predict the concentration of PM2.5/PM10/O3 for 35 stations in Beijing. Please check **sample_submission.csv** for the format of predicting results. Submitted files:

1. Predicting results, named as submission.csv.
2. Report. This report should be at least 5 typed pages, named as report.pdf. Only one report is required for a group, but you need to write the name, student id and task assignment for every

group member. Please note that the report is very important, so it should be written very carefully. This is the most important evaluation for the final score of your project.

3. Code. Put all your codes in a folder named as src. Program entry should be named as main.py

You need to pack them together, named as msbd5002project_groupID.zip.

The deadline is November 30.

Notes

1. You can use any algorithm you learned.
2. Real-world data contains noise, missing values or even mistakes. Data cleaning and pre-processing are necessary.
3. Feature engineering is important, you need to generate features on your own.
4. If your code is very complex, please add readme file.
5. Please email us the name and student id for everyone in your group before **September 30**, and we will announce your group ID after that.
6. **The deadline is November 30.**