Final Project

Tuesday 7th April, 2020

1 Aim

This year's project is the "KDD CUP of Fresh Air" competition. You are going to work with historical air quality data and meteorological (weather) data of two cities, Beijing and London, and the goal is to forecast the air quality in **future**. To do that, you need to predict the concentration of PM2.5, PM10 and O3 for 35 stations in Beijing and PM2.5 and PM10 for 13 stations in London.

2 Data Description

You are going to work using air quality and meteorological data for Beijing and London. The data provided are : a) for Beijing Air Quality data, Observed Meteorology Data and Grid Meteorology Data b) for London : Air Quality data and Grid Meteorology Data.

The data consists of the following files:

- beijing_17_18_meo.csv
- beijing_17_18_aq.csv
- beijing_201802_201803_aq.csv
- beijing_201802_201803_me.csv
- Beijing_AirQuality_Stations.xlsx
- Beijing_historical_meo_grid.csv
- Beijing_grid_weather_station
- London_historical_aqi_forecast_stations.csv
- London_historical_aqi_other_stations.csv
- London_AirQuality_Stations.csv
- London_historical_meo_grid.csv

- London_grid_weather_station

You can find the coordinates of the Air quality stations in the files: Beijing_AirQuality_Stations, London_AirQuality_Stations. In Beijing_grid_weather_station, London_grid_weather_station files you can find the coordinates of the grid weather points (651 points of grid weather data in Beijing and 861 points of grid weather data in London are provided).

For London, information from some extra stations are given concerning the air quality data (London_historical_aqi_other_stations).

More information about the data you can find in the APPENDIX (section 5.2).

3 Model

Before starting your analysis it is very important to understand the data and the objective of this project.
Using the meteorological data you have to predict the air quality of the cities **in the future**. This means that during the test you don't have access to the meteorological data. Based on your training you have to predict the pollution level in two cities, Beijing and London. To predict the pollution level for Beijing you have to predict the concentration of PM2.5, PM10 and O3 for 35 stations in Beijing. You can find the name of the station and their coordinates in Beijing_AirQuality_Stations. The pollution level prediction for London should be made by predicting the concentration of PM2.5 and PM10 for 13 stations (London_AirQuality_Stations). The prediction should be done hourly.

For training and validation use the data until 20 of March 2018 and use the 21^{th} and 22^{th} of March for testing. **It is mandatory to use three or more different algorithms in addition to a baseline.** Split the data set properly in training and validation sets taking into account that **the order matters**. Use cross-validation to select the best model based on **statistical significant test**. Only for the best model you will use the test data (of course you have to include in your report and present all the models that you have tried). Before starting your analysis you have to clean/prepare your data (merge different files, check/ handle missing variables, etc.).

The final model we would expect is a model that can work on universal data, which means it can give a reasonable prediction on all kinds of the products.

You are free to try different approaches and models, you can use ready libraries to test your algorithm. The main focus is to see the model you have came up but also the other approaches that you tried and deep understanding of each one.

4 Report

Your report has to include the approaches that you followed and main results but not only for your final model. We want a full picture of what exactly you have done and how. You should also discuss the different performances you have with your methods and explain why these work or not. What is important is to show us that you have a good understanding of the problem and of how to model it, what are the problems you encountered and how you solved them.

Note that this is an open project, you can try many different approaches as long as they make sense to get the best performance (creative ideas are always welcome).

5 APPENDIX

5.1 Background

Over the past years, air pollution has become progressively more severe in many large cities, such as Beijing. In 2011, an article in the Los Angeles Times cited Dane Westerdahl, an air quality expert from Cornell University, describing the air quality of Beijing as 'downwind from a forest fire'. Among different air pollutants, air particles, or Particulate Matters (PM), are one of the deadliest forms. Particles with a diameter of 2.5 m or less (called PM2.5) can penetrate deeply into human lungs and enter blood vessels, causing DNA mutations, cancer, central neural system damage, and premature death.

Existing biomedical research demonstrates that, once inhaled, PM2.5 can hardly be self-cleaned by the human immune system. Therefore, accurately monitoring and predicting the concentration of PM2.5 and other air particles have become increasingly crucial. With precise predictions of air pollution levels, the public and governments can respond with appropriate decisions, such as closing schools and discouraging outdoor activities, to greatly mitigate the harmful consequences of air pollution.

5.2 Data

The Air Quality Data contains the concentration of several major air pollutants: PM2.5 (ug/m3), PM10 (ug/m3), and NO2 (ug/m3). In addition, the concentrations of CO (mg/m3), O3 (ug/m3) and SO2 (ug/m3) from Beijing are provided.

The Mereological (Weather) Data contains information about the weather, temperature, pressure, humidity, wind_speed and wind_direction.

- **weather** indicates the weather type around a given weather station at a given timestamp. The weather type contains the following possibilities:

  Sunny, Clearm Mostly Sunny, Mostly Clear, Cloudy, Partly Cloudy, Overcast, Showers, Scattered Showers, Light Showers, Heavy Showers, Snow Showers, Light Snow Showers, Fog, Freezing Fog, Sandstorm, Dust, Dust Storm, Sand, Heavy Sandstorm, Haze, Thundershower, Lightning, Thunderstorm, Thundershower with Hail, Hail, Needle Ice, Icy, Sleet, Light Rain, Rain, Heavy Rain, Rainstorm, Heavy Rainstorm, Extreme Rainstorm, Light Snow, Snow, Heavy Snow, Blizzard, Freezing Rain

- **temperature** is measured with a thermometer in a weather station. The temperature data of Beijing is in centigrade scale (°C).

- **pressure**, or atmospheric pressure, is the force per unit area exerted by the weight of the air. The unit is hectopascal or hPa (1 hPa = 100 Pa).

- **humidity** is the measure of the amount of water vapor present in the air. The unit of humidity is the percentage (%).

- **wind_speed**, or speed of the wind, is measured by anemometers in weather stations instrument. The unit of wind speed is meter per second (m/s) or kilometer per hour (km/h).

- **wind_direction**, or wind direction, is the direction from which it originates. For example, a northerly wind blows from the north to the south. Wind direction is measured in degrees clockwise from due north and so a wind coming from the south has a wind direction of 180 degrees; one from the east is 90 degrees, etc. If the wind speed is less than $0.5m/s$ (nearly no wind), the value of the wind_direction is 999017.

Observed Weather Data Vs. Grid Weather Data
Observed weather data is measured by instruments, such as thermometers (for temperature) and anemometer (for wind speed).

Although the observed weather data is the first-hand record of atmospheric conditions, as human errors or equipment faults occur from time to time, a quality control process is required to identify and correct those errors.

Different organizations, such as National Oceanic and Atmospheric Administration (NOAA), use similar, but not identical, methods to process, correct and smooth data. This process takes station observations, satellite image and other meteorology and atmosphere data, conducts complicated calculation and returns a continuous data distribution over earth surface, while observed data can be measured only at the place of a weather station.

651 points of grid weather data in Beijing and 861 points of grid weather data in London are provided, because Air Quality forecasting requires continuous data

on a much larger scale of time and space. In addition, observed weather data from 18 weather stations in Beijing are provided.