Machine Learning

CSCI 5622

Prediction of PM2.5 level at an AQMD station
using neighboring AQMD stations' data

Team Members

Nitish Venkatesh, Septankulam Ramakrishnan

Meghana Vasanth Shettigar

Jooseok Lee

- Team name
  - Astron

- Gr
  - Nitish Venkatesh, Septankulam Ramakrishnan
  - Meghana Vasanth Shettigar
  - Jooseok Lee
- Project Title
  - Prediction of PM2.5 level at AQMD station: using nearby AQMD stations' data
- Motivation
  - Predicting the level of particulate matter (PM2.5) is becoming increasingly crucial in the field of public health. For example, the United States Environmental Protection Agency (US EPA) designates particulate matter as an important source of air pollution (https://www.epa.gov/pm-pollution). However, not every station is able to measure the level of PM2.5. For instance, only 13 of 30 stations in the South Coast Air Quality Management District (AQMD) can monitor the level of PM2.5. We plan to develop a machine learning model that can predict the level of PM2.5 of stations that have no PM2.5 level monitoring capability. To achieve this, we will rely on the fact that the air quality of one station has a spatiotemporal correlation to the air quality of nearby stations [1]. We will try to build a model that utilizes historical data of PM2.5 of nearby stations that have monitoring capability. We will utilize other commonly monitored air pollution data, such as level of NO2 and CO, and basic meteorological data, such as wind direction and speed, to increase the performance of the model.
  - Since spatiotemporal relationships between monitoring stations are extremely complex, we believe machine learning is the right tool to infer the relationships based on huge amounts of data.
- Data / Data Plan
  - AQMD offers historical data of various components, including the level of PM2.5, CO, wind speed, etc. for each monitoring station. 13 out of 30 stations have data on the level of PM2.5, which can be used to build and test models. We can download data for each component and station in a CSV file format.
  - Initially, we plan to obtain data for a period of six months and then plan to extend gradually. There are 4,326 rows for six months of data, meaning at least 4.326 samples for each station.
  - We have full access to the dataset, which can be downloaded.

    - We will only use data from 13 stations that have a level of PM2.5. We will pretend each targeted station does not have the data on PM2.5 and use the other 12 stations' PM2.5 data (along with other data) to predict the target.

- Features
    - Data from the target station
        - Commonly monitored air pollution data (e.g. NO2, CO, O3, etc.)
        - Basic meteorological data (e.g. wind direction, wind speed)
        - We do not use the PM2.5 level of a target station even though it measured the level of PM2.5 to simulate the situation of not having a PM2.5 level monitoring capability
    - Data from nearby stations that have PM2.5 monitoring capability
        - PM2.5 level
        - Commonly monitored air pollution data (e.g. NO2, CO, O3, etc.)
        - Basic meteorological data (e.g. wind direction, wind speed)
        - (If possible) Give weight to data using the distance to the target station or other geographical features
- Target
    - PM2.5 level of the target station
- Metric
    - RMSE
    - MAE
    - Other metrics related to regression

- Reference
    - [1] Xiao, F., Yang, M., Fan, H. et al. An improved deep learning model for predicting daily PM2.5 concentration. Sci Rep 10, 20988 (2020). https://doi.org/10.1038/s41598-020-77757-w