paper05_data_aggregation

October 10, 2021

1 Data Aggregation Across Data Sources

We have 3 different sources of data:

1. Our sensor data: that has the Indoor Air Quality and Indoor Environmental Data.
2. SINAICA: Outdoor Air Quality Monitoring Data from the Government.
3. OpenWeatherData: Outdoor Environmental Data.

We need it to be available that data to the models we plan to train. In the following sections this process is detailed.

```
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
```

```
               CO        NO       NO2       NOx        O3       PM10  \
987      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
988      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
989      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
990      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
991      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
…              …       …       …        …        …        …
1582081  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1582082  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1582083  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1582084  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1582085  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348

             PM2.5       SO2  month  day  hour                      datetime  \
987      22.000000  0.004000      2   12     6 2021-02-12 06:05:35.846304417
988      22.000000  0.004000      2   12     6 2021-02-12 06:05:38.837326527
989      22.000000  0.004000      2   12     6 2021-02-12 06:05:47.812360048
990      22.000000  0.004000      2   12     6 2021-02-12 06:05:50.803695202
991      22.000000  0.004000      2   12     6 2021-02-12 06:05:53.795462847
…                …         …    …   …    …                           …
1582081  16.391304  0.001304      9   18     0 2021-09-18 00:59:47.142104626
1582082  16.391304  0.001304      9   18     0 2021-09-18 00:59:50.136709690
1582083  16.391304  0.001304      9   18     0 2021-09-18 00:59:53.131285429
1582084  16.391304  0.001304      9   18     0 2021-09-18 00:59:56.125959396
1582085  16.391304  0.001304      9   18     0 2021-09-18 00:59:59.120573282

         minute  temperature  pressure  humidity  gasResistance    IAQ
```

```
987        35.0        21.51    777.41     44.04       152149.0   34.7
988        34.0        21.51    777.41     43.98       152841.0   33.6
989        32.0        21.54    777.41     43.73       153259.0   31.5
990        32.0        21.53    777.41     43.70       152841.0   31.5
991        30.0        21.52    777.41     43.70       153399.0   30.2
…         …          …       …       …            …    …
1582081   138.0        26.00    782.92     56.34       916837.0  138.2
1582082   138.0        26.00    782.92     56.33       917462.0  137.7
1582083   138.0        26.00    782.90     56.34       916837.0  137.6
1582084   136.0        26.00    782.92     56.35       921233.0  136.0
1582085   134.0        25.99    782.92     56.35       922497.0  134.5

[1581099 rows x 18 columns]

               CO        NO       NO2       NOx        O3       PM10  \
0        2.500000  0.244000  0.035000  0.205000  0.002000  57.000000
1        2.500000  0.244000  0.035000  0.205000  0.002000  57.000000
2        2.500000  0.244000  0.035000  0.205000  0.002000  57.000000
3        2.500000  0.244000  0.035000  0.205000  0.002000  57.000000
4        2.500000  0.244000  0.035000  0.205000  0.002000  57.000000
…           …       …       …       …       …        …
1605252  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1605253  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1605254  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1605255  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348
1605256  0.765217  0.009174  0.027826  0.039043  0.015304  20.304348

             PM2.5       SO2  month  day  …  pressure_outdoor  \
0        25.000000  0.005000      2   12  …              1020
1        25.000000  0.005000      2   12  …              1020
2        25.000000  0.005000      2   12  …              1020
3        25.000000  0.005000      2   12  …              1020
4        25.000000  0.005000      2   12  …              1020
…            …       …    …  …  …               …
1605252  16.391304  0.001304      9   18  …              1015
1605253  16.391304  0.001304      9   18  …              1015
1605254  16.391304  0.001304      9   18  …              1015
1605255  16.391304  0.001304      9   18  …              1015
1605256  16.391304  0.001304      9   18  …              1015

        humidity_outdoor  wind_speed  wind_deg  rain_1h  rain_3h  clouds_all  \
0                     44        0.00         0      0.0      0.0           1
1                     44        0.00         0      0.0      0.0           1
2                     44        0.00         0      0.0      0.0           1
3                     44        0.00         0      0.0      0.0           1
4                     44        0.00         0      0.0      0.0           1
…                  …         …       …      …      …         …
1605252               93        1.37       199      1.0      0.0           0
1605253               93        1.37       199      1.0      0.0           0
1605254               93        1.37       199      1.0      0.0           0
1605255               93        1.37       199      1.0      0.0           0
1605256               93        1.37       199      1.0      0.0           0

         weather_id  weather_main  year
0               800         Clear  2021
1               800         Clear  2021
2               800         Clear  2021
3               800         Clear  2021
4               800         Clear  2021
```

```
…             …           …   …
1605252         500         Rain  2021
1605253         500         Rain  2021
1605254         500         Rain  2021
1605255         500         Rain  2021
1605256         500         Rain  2021

[1605257 rows x 32 columns]

CO                            float64
NO                            float64
NO2                           float64
NOx                           float64
O3                            float64
PM10                          float64
PM2.5                         float64
SO2                           float64
month                           int64
day                             int64
hour                            int64
datetime               datetime64[ns]
minute                        float64
temperature                   float64
pressure                      float64
humidity                      float64
gasResistance                 float64
IAQ                           float64
temperature_outdoor           float64
feels_like                    float64
temp_min                      float64
temp_max                      float64
pressure_outdoor                int64
humidity_outdoor                int64
wind_speed                    float64
wind_deg                        int64
rain_1h                       float64
rain_3h                       float64
clouds_all                      int64
weather_id                      int64
weather_main                   object
year                            int64
dtype: object
```

1.1 Interpolation of Hourly Data

We found that the dataframe contains repeated records on the columns of hourly data: SINAICA Gov't Air Quality Monitoring and OpenWeatherData.

We think that the repeated data can be an issue, as the data moves very abruptly from a record call it at 10:57 and 11:00.

We propose an approach similar to the imputations using the interpolation incorporating noise, that could avert the overfitting issue on our machine learning and deep learning training.

1.2 Resampling

To reduce training time we propose to have a resampling of the data.

In the following subsections we create those resampled-data dataframes.

1.2.1 1 Minute Resampling

1.2.2 2 Minute Resampling

1.2.3 3 Minute Resampling

1.3 References

- https://scikit-learn.org/stable/modules/linear_model.html#generalized-linear-regression
- https://pythonhealthcare.org/2018/05/03/81-distribution-fitting-to-data/
- https://medium.com/@amirarsalan.rajabi/distribution-fitting-with-python-scipy-bb70a42c0aed
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KernelDensity.html?highlight=kernel%20density#sklearn.neighbors.KernelDensity