2017 Fall Data Mining

# Homework 0 - PM2.5

*Data preprocessing and exploring the dataset*

Due Date: 23:59, October 13 Friday, 2017

TA: 池昊頤 anarchih.tw@gmail.com

In this homework, you need to do some data preprocessing, and then get some basic information about the dataset via tools.

**Dataset:**

1. https://sites.google.com/site/cclljj/dataset-airbox
2. We will use “March 2017, Taiwan” dataset. (1 month)

Columns:

| | | | | |
|---|---|---|---|---|
| Date | Time | device_id | PM2.5 | PM10 |
| PM1 | Temperature | Humidity | lan | lon |

**Preprocessing Tasks:**

1. Discover bad data and remove/fix it.
    a. Remove the anomal records. (Ex: value of PM2.5 is 0)
    b. Remove the sensors with few data or long time gap.

    Decide the conditions by yourself, and remember to write down all conditions you set in the report.

2. Data alignment
    We require aligned-time data for future works. However, the sensors don’t have a static record time. In order to solve this problem, we can use linear interplolation. (or any methods you like)

    For example, if we want to align the sample rate to 1 record every 10 minutes, we can calculate the pm2.5 value at 17:10 by using the value at 17:04 and 17:18.

    Original Data

| time | 17:04 | 17:18 | 17:22 | 17:29 | 17:36 |
|---|---|---|---|---|---|
| pm2.5 | 10 | 20 | 30 | 40 | 50 |

New Data

| time | 17:10 | 17:20 | 17:30 |
|---|---|---|---|
| pm2.5 | 14.286 | 25 | 41.429 |

3. Save new dataset to DB or files.

**Observation Tasks (Visualization):**

1. Compare different sensors in same time interval.
   ex: Sensor 1 on 8/10 v.s. Sensor 2 on 8/10

2. Plot all sensors on the map and describe the dataset.
   If you don't know how to plot points on map. Try this online tool "GPS Visualizer" (http://www.gpsvisualizer.com/map_input?form=data)

3. Try to find some interesting obesrvations.

**Query Tasks:**

1. How many sensors are there in the dataset?
   統計總共有多少個感測器。

2. Which sensor recorded the highest temperature in March? What's the temperature? And where's the sensor?
   3 月份最高的溫度是哪個感測器記錄到的？溫度是幾度？以及該sensor位在哪裡？

3. What were the maximal PM2.5 values of each sensors on 3/5?
   不同感測器在 3/5 的 PM2.5 最大值分別是多少？

4. Try to find some interesting queries.

**Time Series Data Comparsion Tasks:**

Choose two sensors to do the following steps.

Q: Sequence from Sensor 1
C: Sequence from Sensor 2
Distance(Q, C): Distance between Q & C

1. Offset Translation
   Q = Q - mean(Q), C = C - mean(C)

2. Amplitude Scaling
   Q = (Q - mean(Q)) / std(Q), C = (C - mean(C)) / std(C)

3. Linear Trend Removal
   Q = detrend(Q), C = detrend(C)

4. Noise Removal
   Q = smooth(Q), C = smooth(C)

5. Calculate Distance(Q, C) at each step, and compare the difference between original data and transformed data.

**Report:**

1. In the preprocessing part, please remember to write down all conditions you set and some observations (Ex: How many sensors are removed because of few records)

2. For each observation task, please show at least one graph by any tools and explain what you have observed in report.

3. For each query task, if you use some database, please include your SQL in the report. If you write a program or use a library, please hand in with code.

4. Please hand in your report with pdf, odt or html format (ipython notebook is allowed)

5. If you hand in your report with doc(x) or other formats, You may lose some points.

6. If you have any questions or suggestions, feel free to contact me.