2017 Fall Data Mining

Homework 2

Clustering

Due Date: 23:59, December 1st, Friday, 2017

TA: 池昊頤 anarchih.tw@gmail.com

In this homework, the major task is turn data into different clusters, and explain what the cluster means.

Dataset:

1. https://sites.google.com/site/cclljj/dataset-airbox (Same as Hw0)
2. You can use the dataset before/after preprocessing.

Preprocessing:

1. Select all sensors in Taiwan.

Tasks:

1. Spatial Clustering:
    a. Use geometric information to do clustering. (i.e. lat and lon)
    b. Use two kinds of clustering methods and compare the differences.

2. Spatial + PM2.5 Clustering:
    a. Combine geometric information and PM2.5 data to do clustering.
    You can choose any timestamp, Ex: PM2.5 at 2017/3/10 17:10:00
    b. Normalize the data and do clustering again.
    In this case, we can get three-dimension data points (lat, lon, pm2.5). If we use raw values to do clustering, the clustering results may highly depend on pm2.5 because geometric information has smaller difference. Therefore, you should normalize the data and do clustering again. (Try any normalization you like)
    c. Compare the differences between clustering before normalization and clustering after normalization.
    d. Do some comparisons between Task 1 and Task 2.

3. Temporal Clustering:
    a. Use a static time interval to do clustering.
    For example, if the sample rate is 1 data / 10 minutes and we use all data in 2/1, each sensor will have 24 * 6 = 144 points. Therefore, you can consider these points as different dimensions, and then do the 144-dimension clustering.
    b. Use PCA to reduce the dimension, and do clustering again.

c. Compare the differences between clustering before PCA and clustering after PCA.

4. Others:
    a. Try to define clustering features by yourself and do clustering.
    For example, you can combine spatial and temporal information to do clustering.

For each task (except Task 1), you can try only one clustering method, but you should use three kinds of clustering methods for this homework in total.
For example:

1. KMeans
2. Agglomerative Clustering
3. DBSCAN

Reports:
In the report, you should

1. clarify all parameters and methods you use.
2. do some observations and comparisons with proper visualization. (ex: chart, map)
3. explain the results.

Your code should be submitted with the report.