Homework 3

Classification, regression ,and other prediction model

Due Date: 23:59, December 29, Friday, 2017

TA: 蔡子豪 stu8978@gmail.com

In this homework, the major task is prediction, that is what we like most, maybe you can try to predict your final score after this homework:)

There are a lot of prediction models, classification model predict labels, regression model predict real values, and there are some prediction model are used in different types of data.

There is a Hackmd for us to discuss this homework.(except answers)

Dataset:

We'll use data from **201703**~201707 (after preprocessed in HW0)

(You can download raw data from https://www.citibikenyc.com/system-data)

Schema:

The preprocessed data should contains the following

Every station's information

id, name, lat, lng

Every stations' flow data

id, time, in-flow, out-flow

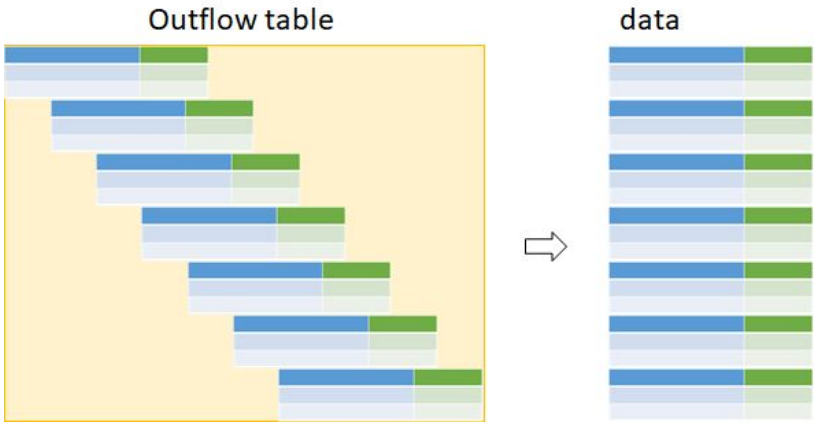
What we want to do is using historical (14 days) data to predict every stations' outflow tomorrow (1 day)

1. Extract following values
   a. station_id
   b. outflow(and this is we want to predict)
2. (When you are think this is a classification problem,you may need to discretize outflow,eg: 26/5=5,34/5=6)
3. We want to use previous (14 days) data to estimate next days' outflow

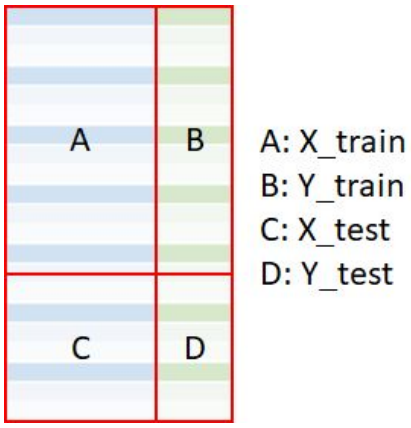
and we can use a sliding window to increase our data(shift **k** days each time, and you can determine the **k**)

Station id	Historical_outflow(14*24)	Target(1*24)
1		
2		
3		

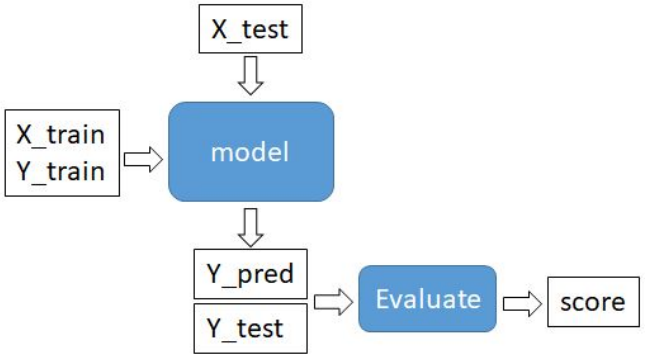
(14*48 and 1*48 ,not *24)



4. **Split the data randomly** to training data and testing data (70% / 30%)



5. Finally we can use these data to train our model.
6. Due to we have 48 target for 48 time slot in single days,you may need to build 48 classifiers, and for each classifier ,it's a multi-class classification problem ,you may want to use one vs one ,or one vs rest strategy.



Tasks:

A:

1. Please try following models (as **classification** problem), compare the computation time and result (**average accuracy for 48 timeslot**).

a. K-Nearest-Neighbor
b. Naive Bayes
c. Random Forest
d. Support vector machine(SVC)
e. other

2. Print the **confusion matrix for predicting the first hour in one day** for Naive Bayes.
3. What is the performance with different parameters in SVM

B.

1. Please try following models (as **regression** problem), compare the computation time and result (**Mean square error**).
   a. ARIMA
   b. Bayesian regression
   c. Decision tree regression
   d. Support vector machine(SVR)
   e. other

C.

1. Try your own method to solve this prediction problem,and give a result and some explanation.
   Some hint:
   You can try other model,use diffenent parameters, or consider different features,like day of week, other nearby stations' inflow or the transition probability.

Report:

In this homework, you should try the previous methods and make some observation, compare different method and parameters with charts,explain the result and see if the output meet your expect.

Your code should be submitted with the report

If you have any questions or suggestions,feel free to contact me:)