Homework 3

Classification, regression, and other prediction model

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

TA: 池昊頤 anarchih.tw@gmail.com

In this homework, the major task is prediction.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.

Dataset:

1. https://sites.google.com/site/cclljj/dataset-airbo
2. You need to use **three-month-data** in this homework. (January, February & March 2017)
3. In this homework, we only need **one** sensor. So please select one sensor to do preprocessing (hw0) and then finish the following tasks.

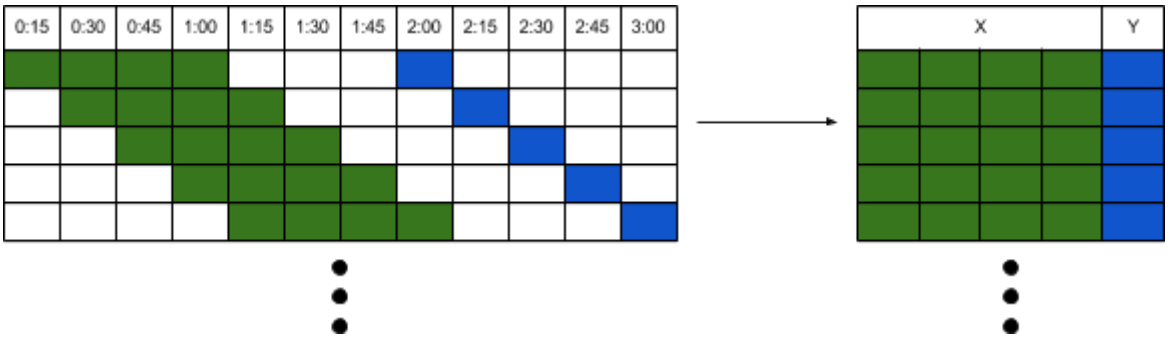
Objective:

Use historical data to predict the PM2.5 value after 1 hour.

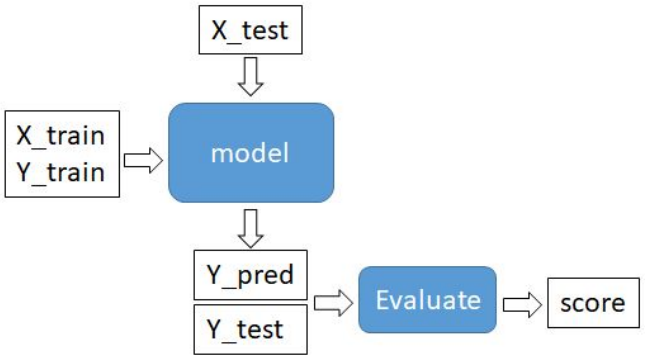
1. Define the input and output (X and Y)
   For example, if the sample rate is 1 data / 15 minutes and we use previous hour to predict the PM2.5 value next hour, the **Input(X)** and **Output(Y)** will look like the following figure. Because the length of dataset is three month, we can approximately get 4 * 24 * 30 * 3 data points.

   In this case, our input size (i.e. number of features) is 4 because we only use one-hour-data to predict PM2.5 values, but you can try to use N-hour-data to do the prediction. Besides, you can also use not only PM2.5 but also PM10, humidity and temperature to do the PM2.5 prediction.

   Attention: You are not allowed to change the output. That is, you need to predict the PM2.5 value after one hour, not two hours, three hours ...etc.



4. Randomly split the data to training data and test data. (70% / 30%)
5. Train models and do evaluations.



Tasks:

1. Try classification models to predict PM2.5 value:
    a. Descritization
    Because the output of classification models are categories (labels), you cannot apply the models to your data directly. To solve this problem, you should encode the output(Y) to categories (labels) by using discretization methods before classification.
    b. Try following models
        i. K-Nearest-Neighbor
        ii. Naive Bayes
        iii. Random Forest
        iv. Support vector machine (SVC)
        v. Others (ex: Neural Network)
    c. Evaluation and Comprasion

2. Try regression models to predict PM2.5 value:
    a. Try following models
        i. Bayesian Regression
        ii. Decision tree Regression
        iii. Support vector machine (SVR)
        iv. Others (ex: ARIMA)
    b. Evaluation and Comprasion

3. Try to define a classification (or regression) problem and solve it:
    a. Clarify the problem you define.
    b. Use at least two methods (i.e. two models) to solve the problem.
    c. Evaluation and Comprasion

Report:
For each Task, you should clarify

1. input / output definitions
2. all parameters you use
3. label definitions (classification problem only)
4. evaluation methods you use
5. evaluation results and comprasions (including computation time)

Try to use different label definitions, input/output definitions, algorithms and parameters to do the prediction. And remember to evaluate the models and compare the results.

Your code should be submitted with the report

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