機器學習概論 Introduction to Machine Learning
Assignment 1 – Dengue Case Prediction
Department: 理學院學士班
Student ID: 108020017
Name: 黃珮綺
-

1. Regression Model for Basic Part

The regression model I use in basic part is:

$$Y_t = w_0 + w_1 * X_t + w_2 * Y_{t-1} + w_3 * Y_{t-2}$$

The model predicts the number of weekly dengue cases based on temperature, the number of cases in the past two weeks. Furthermore, the models for the three cities are generated individually here.

2. Regression Model for Advanced Part

In this part, I generate only one model that predict the number of dengue cases for all cities. The variables are those in basic part plus the precipitation rate of each weeks. The additional variables to be used are selected by the program, which best fit the model (depending on MAPE). Hence, the regression model in advanced part is:

$$Y_t = w_0 + w_1 * X_t + w_2 * Y_{t-1} + w_3 * Y_{t-2}$$
$$+w_4 * Precipitation + w_5 * Population + w_6 * Age0 - 4(\%)$$
$$+w_7 * PeoplewithDisabilities(\%) + w_8 * Unemployedpolpulation(\%)$$

3. Difficulty I encountered

I find it hard to identify outliers in the given data, since the distribution of the temperature clusters between 20°C to 30°C, and the relation between temperature and the number of dengue cases seems to be at random. Furthermore, the case numbers of City A and City B are quite high in the first 20 weeks. It is difficult to tell whether they describe the regression model well or not.

4. Solve the difficulty and reflections

I plot out the distribution of temperatures of each city, observing data that seems unreasonable. More specifically, temperatures above 40°C are impossible, hence, weeks with high temperature are consider as outlier. Those weeks with abrupt changes in temperature are also included. I manually classify out these data, and replace them with the average temperature of the past and the next week (taking the continuity of the temperature into account). Some special cases are also handled at the same time.

In this way, the efficiency and the accuracy of training the model is low, since I should repeatedly check if I was deleting a “correct” data. Statistical methods should be imported to help removing outlier, which is more efficient. What I already know is computing jackknife residuals to identify outliers. Though, it is quite complicated if implemented in mathematical ways.

At last, I choose the case numbers of the past two weeks as variables in both models. These variables will explain the trend of the number of cases. However, more checks such as whether it is overfitting is still needed to be implemented.