Principles and Applications of Data Science
Homework #3

Due: June 5, 2019

This assignment is to practice how to generate a linear regression from a numeric dataset. Here, we provide the PM 2.5 data file (.csv) for practice. Please download the dataset from the :URL, https://opendata.epa.gov.tw/ws/Data/ATM00756/?$format=csv. In the dataset, there are fifty nine **attributes**; however, we only interest the **eleven** attributes for exploring the linear regression. These attributes are **Na_ion, K_ion, Mg_ion, Ca_ion, SO4_ion, NH4_ion, NO3_ion, Cl_ion, OC, EC and PM2.5**. The meanings of the above eleven attributes are sodium ion, potassium ion, magnesium ions, calcium ion, sulfate ion, ammonium ion, nitrate ion, chloride, organic carbon, elemental carbon and PM2.5 mass concentration, respectively. The first ten columns are independent variables and the last one is the dependent variable. Assume the linear regression can be denoted as $y = c + w_1x_1 + w_2x_2 + ... + w_{10}x_{10}$ where y is the dependent variable, x_i are independent variables, c is the constant and w_is are the coefficients of the linear regression. Please show the coefficients of the linear regression in order (i.e., c, w_1, w_2,..., w_{10}) with the following situation:

1. Calculating the linear regression from the raw data directly. (You can choose one of the approaches in class for implementation; of course, you must make sure that you won't get a singular matrix if you use the matrix approach.)

2. Improving the linear regression from question 1 and get a new linear regression if the coefficients in question 1 are meaningless.

The example will be put on the *ischool* (`http://www.ischool.ntut.edu.tw/`) platform of school for you to access.

Homework Submission

- Please compress and upload your homework file named as `HW3_LR_studentID.zip( or .rar)` to *ischool* platform (`http://www.ischool.ntut.edu.tw/`). The submitted file includes:
  - The source code (.ipynb, .py, or others).
  - A text file (.txt, doc, or docx). In this file, you must show the coefficients matrix of the linear regressions in questions 1 and 2.
- The **deadline** is the **midnight of June 5**, 2019 and **Late work** is not acceptable.
- Honest Policy: We encourage students to discuss their work with the peer. However, each student should write the program or the problem solutions on her/his own. Those who copy others work will get 0 on the homework grade.