# 2019 ML HW1 Report

**學號：B05705042　系級：資管四　姓名：皇甫立翔**

## 1. 記錄誤差值 (RMSE)，討論兩種 feature 的影響。

第一種model(所有污染源當作feature)的RMSE = 5.44384(public) + 5.48710(private) = 10.93094
第二種model(只取pm2.5當作feature)的RMSE = 5.95141(public) + 5.78967(private) = 11.74108

從結果來觀察，可以發現只取pm2.5當作feature的結果明顯比較差勁。大二的時候統計學我也做過pm2.5相關的期末報告研究，我認為這是因為pm2.5與其他天氣因子息息相關的緣故，像是pm2.5本身也算是pm10的一部份，他們之間的關聯性就非常的大，再者像是溫度、空氣濕度、風力等等天氣因子也在一定程度上影響著pm2.5濃度。這些變化因子彼此會交互影響，所以如果將相關的影響因素也加進feature來參考，所預測出來的pm2.5濃度也會比較精準。

## 2. 解釋什麼樣的 data preprocessing 可以 improve 你的 training / testing accuracy，ex. 你怎麼挑掉你覺得不適合的 data points。請提供數據 (RMSE) 以佐證你的想法。

(1) 在一開始，如果只是單純將資料清理乾淨(NR變成0，去掉奇怪符號等等)，並將pm2.5濃度小於2或大於100的資料點踢除來做training的話，得到的分數是5.61595(public)。

(2) 接著再精進踢除資料點的界線，將training data踢除的標準 — pm2.5「小於2或大於100」的界線改為ㄉpm2.5「小於0或大於平均值+3.8*標準差」。同時將testing data當中的pm2.5「小於0或大於平均值+3.8*標準差」的部分，通通assign另一個值：沒有超出範圍的pm2.5資料點的平均。如此得到的分數是5.51889(public)。

(3) 之後再將training data與testing data分別對pm10也做(2)的處理(training的部分不踢掉，改成跟testing一樣做re-assign)，得到的分數為5.45246(public)。

(4) 再將training data與testing data分別對「除了溫度、pm2.5、pm10以外」的所有feature做(2)的處理(training改成re-assign)，得到的分數為7.03363(public)。

到了這邊發現這樣做了太多的人工介入了，最後決定將training data只做pm2.5與pm10的處理，將testing data做除了溫度以外的所有feature的處理。最後得到的分數為5.44384(public)與5.48710(private)。

[備註] 因為用jupyter寫，有一個5.43387的分數不知道是哪裡生出來的，也嘗試過了但是無法再reproduce。

# 3. Refer to math problem

## 1. Closed-Form Linear Regression Solution

### 1-(a).

According to linear regression formula via least square method, we can get that：

$b = \overline{y} - w^T \overline{x}$

$w = \dfrac{\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$

Then applying dataset from the topic:

$b = 3.36 - 1.05 * 3 = 0.21$

$w = \dfrac{4.32 + 0.96 + 0 + 0.74 + 4.48}{4 + 1 + 0 + 1 + 4} = \dfrac{10.5}{10} = 1.05$

### 1-(b).

$$Minimize \quad L_{ssq}(w, b) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))^2$$

**About b :**

$$\frac{\partial Loss}{\partial b} = \frac{-2}{2N} \sum_{i=1}^{N} y_i - (w^T x_i + b) = 0$$

$$\Rightarrow \; b = \frac{1}{N} \sum_{i=1}^{N} (y_i - w^T x_i)$$

$$\Rightarrow \; b = \overline{y} - w^T \overline{x}$$

**About w :**

$$\frac{\partial Loss}{\partial w} = \frac{-2}{2N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))x_i = 0$$

$$\Rightarrow \; \sum_{i=1}^{N} (y_i - (w^T x_i + \overline{y} - w^T \overline{x}))x_i = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i x_i - \overline{y}\sum_{i=1}^{N} x_i - w^T \sum_{i=1}^{N} (x_i - \overline{x})x_i = 0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{N} (y_i - \overline{y})x_i}{\sum_{i=1}^{N} (x_i - \overline{x})x_i}$$

$$\Rightarrow w = \frac{\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

**1–(c).**

$$Minimize \quad L_{req}(w, b) = \frac{1}{2N}\sum_{i=1}^{N} (y_i - (w^T x_i + b))^2 + \frac{\lambda}{2}\|w\|^2$$

**About b : Same as 1–(b).**

**About w :**

$$\frac{\partial Loss}{\partial w} = \frac{-1}{N}\sum_{i=1}^{N} (y_i - w^T x_i - \overline{y} + w^T \overline{x})x_i + \lambda w = 0$$

$$\Rightarrow -\sum_{i=1}^{N} y_i x_i + \sum_{i=1}^{N} \overline{y} x_i + w^T (\sum_{i=1}^{N} (x_i - \overline{x})x_i + \lambda N) = 0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{N} (y_i - \overline{y})x_i}{\sum_{i=1}^{N} (x_i - \overline{x})x_i + \lambda N}$$

$$\Rightarrow w = \frac{\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2 + \lambda N}$$

## 2. Noise and Regulation

$$\mathbb{E}\left[ \frac{1}{2N}\sum_{i=1}^{N} (f(x_i + n_i) - y_i)^2 \right]$$

$$= \mathbb{E}\left[ \frac{1}{2N}\sum_{i=1}^{N} (w^T x_i + w^T n_i + b - y_i)^2 \right]$$

$$= \ \mathbb{E} \ [ \ \frac{1}{2N} \sum_{i=1}^{N} (f(x_i) - y_i + w^T n_i)^2 \ ]$$

$$= \ \mathbb{E} \ [ \ \frac{1}{2N} ( \ \sum_{i=1}^{N} (f(x_i) - y_i)^2 + 2 \sum_{i=1}^{N} (w^T n_i)(f(x_i) - y_i) + \sum_{i=1}^{N} (w^T n_i)^2 \ ) \ ]$$

$$= \ \frac{1}{2N} ( \ \sum_{i=1}^{N} (f(x_i) - y_i)^2 + 0 + \|w\|^2 N \sigma^2 \ )$$

$$= \ \frac{1}{2N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \frac{\sigma^2}{2} \|w\|^2$$

Thus we can find that the input noise would be equal to the addition of a $L^2$ regularization term on the weights.

## 3. Kaggle Hacker

### 3-(a).

$$s_k = \frac{1}{N} (g_x(x_i))^2$$

$$e_0 = \frac{1}{N} \sum_{i=1}^{N} (g_0(x_i) - y_i)^2 = \frac{1}{N} \sum_{i=1}^{N} y_i^2$$

$$Ne_k = \sum_{i=1}^{N} (g_k(x_i) - y_i)^2$$

$$= \ \sum_{i=1}^{N} (g_k(x_i))^2 - 2 \sum_{i=1}^{N} g_k(x_i) y_i + \sum_{i=1}^{N} y_i^2$$

$$= \ Ns_k - 2 \sum_{i=1}^{N} g_k(x_i) y_i + Ne_0$$

$$\sum_{i=1}^{N} g_k(x_i) y_i = \frac{N}{2} (s_k + e_0 - e_k)$$

3-(b).

$$Minimize \;\; a_1 \dots a_k L_{test}(\sum_{k=1}^{K} a_k g_k)$$

$$\Rightarrow \; min \; \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(a_k g_k(x_i) - y_i)^2$$

$$\Rightarrow \; min \; \frac{1}{N}\sum_{k=1}^{K}[\;(\sum_{i=1}^{N} a_k g_k(x_i))^2 - (2\sum_{i=1}^{N} a_k g_k(x_i)y_i) + (\sum_{i=1}^{N} y_i^2)\;]$$

$$\Rightarrow \; min \; \frac{1}{N}\sum_{k=1}^{K}(Na_k^2 s_k - 2a_k\frac{N}{2}(s_k + e_0 - e_k) + Ne_0)$$

$$\Rightarrow \; min \; \sum_{k=1}^{K}(a_k s_k - e_0)(a_k - 1) + a_k e_k$$

$$Let \;\; loss = \sum_{k=1}^{K}(a_k s_k - e_0)(a_k - 1) + a_k e_k$$

Minimize loss and obtain optimal $a_k$ :

$$\Rightarrow \; \frac{\partial loss}{\partial a_k} = \sum_{k=1}^{K} s_k(a_k - 1) + (a_k s_k - e_0)1 + e_k = 0$$

$$\Rightarrow \; \sum_{k=1}^{K} s_k a_k - s_k + a_k s_k - e_0 + e_k = 0$$

$$\Rightarrow \; \sum_{k=1}^{K} 2a_k s_k - s_k + e_k - e_0 = 0$$

$$\Rightarrow \; a_k = \frac{\sum_{k=1}^{K} s_k + Ke_0 - \sum_{k=1}^{K} e_k}{2\sum_{k=1}^{K} s_k}$$