ML Fall 2019 HW1　　學號：B04303006 系級： 經濟四 姓名：劉傳筠

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (2) 題：

(1) 抽全部 9 小時內的污染源 feature 當作一次項(加 bias)

(2) 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註：

a. NR 請皆設為 0，其他的非數值(特殊字元)可以自己判斷

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

c. 第 1-2 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好，kaggle 死線之後便可以無限上傳。

e. 根據助教時間的公式表示，(1) 代表 p = 9x18+1 而(2) 代表 p = 9*1+1

1. (1%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

*(a)用 50%的 training data(shuffled), adagrad, initial learning rate: 2, update: 2000 times

	Private	Public
Model 1	5.62111	5.72449
Model 2	5.91625	6.01190

*(b): 同(a), 但改用全部的 training data (yr1+yr2)

	Private	Public
Model 1	5.54947	5.74641
Model 2	5.88820	5.99213

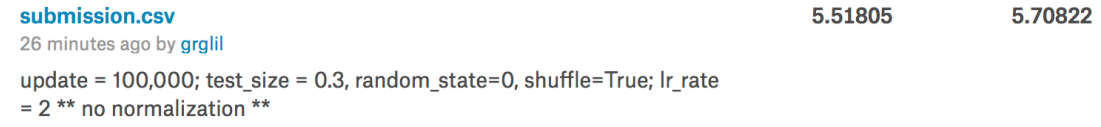
*(c): 同(b), 但 update times * 10 (20,000 次)

	Private	Public
Model 1	5.56805	5.71760
Model 2	5.88820	5.99213

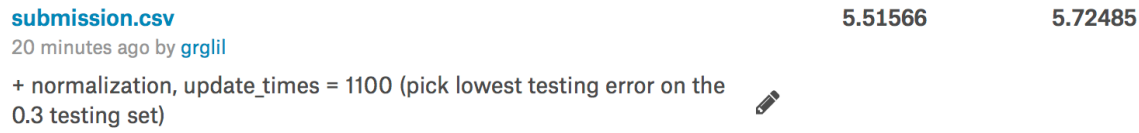
第二種 model 因為只有用前九小時 pm2.5 作為 feature，經過觀察，發現不管在 private/public 的 rmse 都較高，可能是因為其他 17 種 features 也對 pm2.5 有預測的能力，因為第二種 model 沒有包含到，故預測能力較差。

2. (1%)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。

前面先用助教的方式處理奇怪符號和空值。

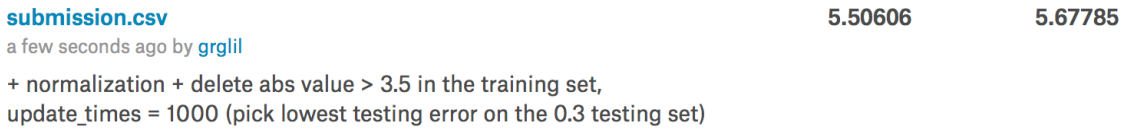


(1) Normalization: 在處理 training data 的時候將其標準化，並存下每個 feature 的 mean and std, 在 testing 的時候就以這兩個為基準去標準化 testing data set。



⇨ 雖然 private 下降但 public 上升了

(2) 去除掉極端值: 將標準化過後還大於 3.5 的值刪掉(原本想用 1.96 但效果不盡理想)



⇨ 比還沒標準化過後的 private 和 public 低(一點點 XD)

3.(3%) Refer to math problem
https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view
(見下一頁)

ML HW1

◆ date/ 10/7 ◆ page/ 1

1. (a) $L_{ssq}(w,b) = \frac{1}{10}\sum_{i=1}^{5}[y_i - (w\cdot x_i + b)]^2$

$$\frac{\partial L}{\partial w} = \frac{1}{10}\cdot 2\sum_{i=1}^{5}[y_i - (w\cdot x_i + b)]\cdot x_i = 0$$

$$\frac{\partial L}{\partial b} = \frac{1}{10}\cdot 2\cdot\sum_{i=1}^{5}[y_i - (w\cdot x_i + b)] = 0$$

$$\sum_{i=1}^{5} y_i - w\sum_{i=1}^{5} x_i + 5b = 0$$

$$\begin{cases} 16.8 - 15w + 5b = 0 \\ 60.9 - 55w - 15b = 0. \end{cases} \times 3$$

$$50.4 - 45w + 15b = 0$$

$$111.3 - 100w = 0 \rightarrow w = 1.113 \rightarrow b = -0.02$$

$$(w,b) = (1.13, -0.02)\ \#$$

(b) Let $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ $\tilde{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \\ b \end{bmatrix}$ $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & 1 \\ x_{21} & & & & 1 \\ x_{31} & & & & \vdots \\ x_{N1} & & & x_{Nk} & 1 \end{bmatrix}$

$$\min L(\tilde{w}) = \min_{\tilde{w}} \frac{1}{2N}\|y - X\cdot\tilde{w}\|^2$$

$$(y - X\cdot\tilde{w})^T(y - X\cdot\tilde{w})$$

$$= y^Ty - 2\tilde{w}^TX^Ty + \tilde{w}^TX^TX\tilde{w}$$

$$\frac{\partial L}{\partial \tilde{w}} = 0 \Rightarrow -2X^T\cdot y + 2X^TX\cdot\tilde{w} = 0$$

$$X^T\cdot y = X^T\cdot X\cdot\tilde{w}$$

$$\therefore \tilde{w} = (X^TX)^{-1}X^Ty \qquad \begin{bmatrix} w \\ b \end{bmatrix} = \tilde{w}\ \#$$

(c) 延用 1.(b) 的 Notation y, $\tilde{w}$, X

New loss $L(\tilde{w})$

$$\min L(\tilde{w}) = \left[(y - X\tilde{w})^T(y - X\tilde{w}) + \frac{\lambda}{2}\tilde{w}^T\cdot\tilde{w}\right]\frac{1}{2N}$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow -2X^Ty + 2X^T\cdot X\cdot\tilde{w} + \lambda\tilde{w} = 0$$

$$\left(X^TX + \frac{\lambda}{2}\cdot I\right)\cdot\tilde{w} = X^Ty$$

$$\tilde{w} = \left(X^TX + \frac{\lambda}{2}\cdot I\right)^{-1}X^Ty \qquad \begin{bmatrix} w \\ b \end{bmatrix} = \tilde{w}\ \#$$

2. $f_{w,b}(x) = w^T x + b$.

$$\tilde{L}_{ssq}(w,b) = E\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i+\eta_i) - y_i\right)^2\right]$$

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

$$= E\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i) - y_i\right)^2\right] + \underbrace{E\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i) - y_i\right)\cdot w^T\eta_i\right]}_{①}$$

$$+ \underbrace{E\left[\frac{1}{2N}\sum_{i=1}^{N} w^T\eta_i\,\eta_i^T w\right]}_{②}$$

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

($f_{w,b}(x_i) - y_i$: 0, $E(\eta_i)$: 0)

② $= \frac{1}{2N}\sum_{i=1}^{N} E\left[w^T\eta_i\,\eta_i^T w\right]$

$= \frac{1}{2N}\sum_{i=1}^{N} E\left[w^T\cdot\sigma^2 I\cdot w\right]$

$= \frac{1}{2N}\cdot\sigma^2\sum_{i=1}^{N} E\left[w^T\cdot w\right]$

$= \frac{1}{2N}\sigma^2\cdot\sum_{i=1}^{N}\|w\|^2$

$= \frac{1}{2N}\sigma^2\cdot N\cdot\|w\|^2$

$= \frac{\sigma^2}{2}\|w\|^2$

$\eta_i\eta_i^T = \begin{bmatrix}\eta_{i1}\\ \eta_{i2}\\ \\ \eta_{ik}\end{bmatrix}\begin{bmatrix}\eta_{i1} & \eta_{i2} & \cdots\end{bmatrix}$

$= \begin{bmatrix}\sigma^2 & 0 & 0\\ 0 & \sigma^2 & \\ 0 & & \sigma^2\end{bmatrix}$

$= \sigma^2\cdot I$

$\therefore \hat{L}_{ssq}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i) - y_i\right)^2 + \frac{\sigma^2}{2}\|w\|^2$ #

3. (a) $e_k = \frac{1}{N}\sum\left(g_k(x_i) - y_i\right)^2 = \underbrace{\frac{1}{N}\sum\left(g_k(x_i)\right)^2}_{S_k} - \frac{1}{N}\cdot 2\sum g_k(x_i)\cdot y_i + \underbrace{\frac{1}{N}\sum y_i^2}_{e_0}$

$\therefore \frac{2}{N}\sum g_k(x_i)\,y_i = S_k + e_0 - e_k$

$\therefore \sum g_k(x_i)\cdot y_i = \frac{N}{2}\left(S_k + e_0 - e_k\right)$ #

Denote $G \in \mathbb{R}^{N \times k}$, $\alpha \in \mathbb{R}^k$, $y \in \mathbb{R}^N$

◆ date/ 10/8 ◆ page/3

(b).

$$L = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_k(x_1) \\ g_1(x_2) & \ddots & & \\ \vdots & & \ddots & \\ g_1(x_N) & & & g_k(x_N) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$L = \frac{1}{N} \| G \cdot \alpha - y \|^2$$

$$\frac{\partial L}{\partial \alpha} = 0 \Rightarrow G^T G \alpha - 2 G^T y = 0.$$

$$G^T G \alpha = 2 G^T y$$

$$\alpha = (G^T G)^{-1} G^T y \quad \text{※}$$

$$(G\alpha - y)^T (G\alpha - y)$$

$$= \alpha^T G^T G \alpha - y^T G \alpha - \alpha^T G^T y - y^T y$$

$$= \alpha^T G^T G \alpha - 2 \alpha^T G^T y - y^T y$$