ML Homework #1 學號：B0902120 系級： 資工四 姓名：曾鈺婷

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

我採用了兩種 gradient descent 的方法，分別是 Adagrad 跟 Adam 兩種不同的演算法，他們各自的更新過程如下：
Adagrad 由前面所有梯度值的平方和來調整 learning rate，如此一來，可以使得前期梯度較小的時候，有較大的 learning rate，而後期梯度較大的時候，可以約束 learning rate。

$$a_t = \sum_{k=1}^{t} (\frac{\partial L_k}{\partial W_k})^2$$

$$W \leftarrow W - \eta \frac{1}{\sqrt{a_t + \epsilon}}$$

Adam 則可以視為更進一步的優化，同時採用 momentum 及 Adagrad 的概念，並且對兩個變數做偏離校正，使得每次的 learning rate 都有一個固定的範圍，讓更新較為平穩。

$$m_t = \beta_1 \, m_{t-1} + (1 - \beta_1) \frac{\partial L_t}{\partial W_t}, \quad \hat{m}_t = \frac{m_t}{1 - \beta_1}$$

$$v_t = \beta_1 \, m_{t-1} + (1 - \beta_2) (\frac{\partial L_t}{\partial W_t})^2, \; \hat{v}_t = \frac{v_t}{1 - \beta_2}$$

$$W \leftarrow W - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

值得一提的是，為了節省運算的時間，我並沒有把整個資料一起丟進去，而是事先 shuffle 資料，隨機抽樣某幾份，這樣既可以節省運算資源，也可以得到類似的效果。

(1) 抽全部 9 小時內的污染源 feature 當作一次項
把每9小時內的 feature 全部節錄，並壓成一維的 vector 當成現在的 feature（ 9 * 18 項）

(2) 抽全部 9 小時內 pm2.5 的一次項當作 feature
把每9小時內的 pm2.5 節錄，當成現在的 feature（ 9 * 1 項）

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

Epoch	Method I (Adam + Adagrad)		Method II (Adam + Adagrad)	
1000	5.49401	6.72330	7.76970	8.14490
2000	5.55162	6.36269	7.76031	8.12057
5000	5.61219	5.82357	7.63874	8.07930
10000	5.67102	5.50867	7.61477	8.01047

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。

首先透過 np.correlation 去比對各項資料對於 pm2.5 的相關性，發現其實他的相關性都蠻低的（小於 0.001），只有第 7 ~ 10 項是比較 ok 的，嘗試只取這四組資料作為 feature 再與上一題的結果做比較，發現雖然有比只取 pm2.5 來得好一些，但效果也還是不如全取，因此最後採用取全部 feature 的方式。另外，在 data 裡面會有一些空格，本來採取的方式是透過內插法補齊，這裡也與直接填入 0 做比較，發現兩者效果其實差不多。所以有刪除的 data 就只有 pm2.5 處於極值的部分，也就是小於 3 或大於 100 者。

Epoch	[7, 8, 9, 10] (Adam + Adagrad)		Fill 0 (Adam + Adagrad)	
1000	6.25430	7.01079	5.55252	6.67961
2000	6.11021	6.85377	5.68000	6.32826
5000	6.10487	6.62066	5.60410	5.79660
10000	6.05211	6.52146	5.55059	5.49355

3. 手寫題

1. Closed-Form Linear Regression Solution - (a)

要使得 $L_{ssq} = \frac{1}{10}\sum_{m=1}^{5}[(w^T x_i + b) - y_i]^2$ 最小，因此我們分別對 w 和 b 做微分。

$$L_{ssq} = \frac{1}{10}\sum_{m=1}^{5}[(w^T x_i + b) - y_i]^2 = \frac{1}{10}\sum_{m=1}^{5}[(y_i - b)^2 + w^2 x_i^2 - 2w(y_i - b)x_i]$$

$$\frac{\partial L_{ssq}}{\partial w} = \frac{1}{10}\frac{\partial \sum_{m=1}^{5}[(y_i - b)^2 + w^2 x_i^2 - 2w(y_i - b)x_i]}{\partial w} = \frac{1}{10}\sum_{m=1}^{5}[2wx_i^2 - 2(y_i - b)x_i]$$

$$= \frac{1}{5}[w\sum_{m=1}^{5}x_i^2 - \sum_{m=1}^{5}x_i y_i + b\sum_{m=1}^{5}x_i] = \frac{1}{5}[55w - 60.9 + 15b] = 11w + 3b - 12.18$$

$$L_{ssq} = \frac{1}{10}\sum_{m=1}^{5}[(w^T x_i + b) - y_i]^2 = \frac{1}{10}\sum_{m=1}^{5}[(y_i - wx_i)^2 + b^2 - 2b(y_i - wx_i)]$$

$$\frac{\partial L_{ssq}}{\partial b} = \frac{1}{10}\frac{\partial \sum_{m=1}^{5}[(y_i - wx_i)^2 + b^2 - 2b(y_i - wx_i)]}{\partial b} = \frac{1}{10}\sum_{m=1}^{5}[2b - 2(y_i - wx_i)]$$

$$= \frac{1}{5}[5b - \sum_{m=1}^{5}y_i + w\sum_{m=1}^{5}x_i] = \frac{1}{5}[5b - 16.8 + 15w] = 3w + b - 3.36$$

兩者微分為 0 則有 L_{ssq} 的最小值，可知 $w = 1.05$ 而 $b = 0.21$

1. Closed-Form Linear Regression Solution - (b)

假設 X_i 為 $[1, x_i]$（$N*(k+1)$ 項），而 W 為 $[b, w]$（$k+1$ 項），則所求為

$$L_{ssq} = \frac{1}{2N}[XW^T - y]^2 = \frac{1}{2N}[WX^TXW^T - 2W^TX^Ty + y^Ty]$$

$$\frac{\partial L_{ssq}}{\partial W} = \frac{1}{2N}\frac{\partial[WX^TXW^T - 2WX^Ty + y^Ty]}{\partial W} = \frac{1}{N}[X^TXW^T - X^Ty]$$

微分之後若得 0 則有 L_{ssq} 的最小值，此時 $X^TXW^T = X^Ty$

若 X^TX 為 invertible 則 $W^T = (X^TX)^{-1}X^Ty$

若 X^TX 不為 invertible 則 $W^T = (X^TX)^{-1}X^Ty = X^+y$ 加入 pseudo inverse 的概念是最佳解

1. Closed-Form Linear Regression Solution - (c)

   假設 X_i 為 $[1, x_i]$（$N*(k+1)$ 項），而 W 為 $[b, w]$（$k+1$ 項），則所求為

$$L_{reg} = \frac{1}{2N}[XW^T - y]^2 + [\frac{\lambda}{2}WW^T - b^2] = \frac{1}{2N}[WX^TXW^T - 2W^TX^Ty + y^Ty] + [\frac{\lambda}{2}WW^T - b^2]$$

$$\frac{\partial L_{reg}}{\partial W} = \frac{1}{2N}\frac{\partial[WX^TXW^T - 2WX^Ty + y^Ty]}{\partial W} + \frac{\lambda}{2}\frac{\partial WW^T - b^2}{\partial W}$$

$$= \frac{1}{N}[X^TXW^T - X^Ty] + \lambda W^T$$

   因為 $X^TX + N\lambda I$ 為 invertible 故 $W^T = (X^TX + N\lambda I)^{-1}X^Ty$

2. Noise and Regulation

$$f_{w,b}(x_i + \eta_i) = w^T(x_i + \eta_i) + b$$
$$= w^Tx_i + w^T\eta_i + b$$
$$= f_{w,b}(x_i) + w^T\eta_i$$

$$\tilde{L_{ssq}} = \mathbb{E}[\frac{1}{2N}\sum_{i=1}^{N}[f_{w,b}(x_i + \eta_i) - y]^2]$$
$$= \mathbb{E}[\frac{1}{2N}\sum_{i=1}^{N}[f_{w,b}(x_i) + w^T\eta_i - y]^2]$$
$$= \frac{1}{2N}\sum_{i=1}^{N}[[f_{w,b}(x_i) - y]^2 + 2(f_{w,b}(x_i) - y)w^T\mathbb{E}[\eta_i] + w^Tw\mathbb{E}[\eta_i^2]]$$
$$= \frac{1}{2N}\sum_{i=1}^{N}[[f_{w,b}(x_i) - y]^2 + 0 + w^Tw * N\sigma^2]]$$
$$= \frac{1}{2N}\sum_{i=1}^{N}[f_{w,b}(x_i) - y]^2 + \frac{\sigma^2}{2}\|w\|$$

3. Kaggle Hacker - (a)

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

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

3. Kaggle Hacker - (b)

假設 X_i 為 $[g_1(x_i), g_2(x_i), \ldots, g_K(x_i)]$（$N * k$ 項），而 W 為 $[\alpha_1, \alpha_2, \ldots, \alpha_K]$（$k$ 項）、

$$\min L_{test} = \min \frac{1}{N}\sum_{i=1}^{N}[\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i]^2 = \min \frac{1}{N}[XW^T - y_i]^2$$

假設 X^TX 為 invertible 則由第一題可知 $W^T = (X^TX)^{-1}X^Ty$，即

$$\left(\begin{bmatrix} g_1(X_1)\ g_1(X_2)\ \ldots\ g_1(X_N) \\ g_2(X_1)\ g_2(X_2)\ \ldots\ g_2(X_N) \\ \\ g_K(X_1)\ g_K(X_2)\ \ldots\ g_K(X_N) \end{bmatrix} \begin{bmatrix} g_1(X_1)\ g_2(X_1)\ \ldots\ g_K(X_1) \\ g_1(X_2)\ g_2(X_2)\ \ldots\ g_K(X_2) \\ \\ g_1(X_N)\ g_2(X_N)\ \ldots\ g_K(X_N) \end{bmatrix}\right)^{-1} \begin{bmatrix} \frac{N}{2}[s_1 - e_1 + e_0] \\ \frac{N}{2}[s_2 - e_2 + e_0] \\ \\ \frac{N}{2}[s_K - e_K + e_0] \end{bmatrix}$$