機器學習 Homework1 Report

學號：B03902013 系級： 資工四 姓名：吳克駿

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (3) 題：

(1) 抽全部 9 小時內的污染源 feature 的一次項(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，其他的數值不要做任何更動

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

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

每個月有 480 個小時，每 10 個小時為一組，總共可以獲得 471 個 data

(1)抽所有汙染源 public:7.46469 private:5.46046 **ans:6.53964**

(2)抽 pm2.5 一次項 public:7.44013 private:5.62719 **ans:6.59624**

兩種在 public 的差距不大，但單純 pm2.5 在 training 時收斂的較快，因此適合在訓練初期做多次的嘗試。而抽取所有汙染源在 private 的成績較好，可能是因為單從 pm2.5 的值來預測，忽略了太多外在的其他因素，而 18 個汙染源中應有幾項與 pm2.5 的值應有絕對的影響，使得預估出來的值較為精準。

2. 將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

抽前 5 小時，可以將每個月的 training data 增加到 475 個

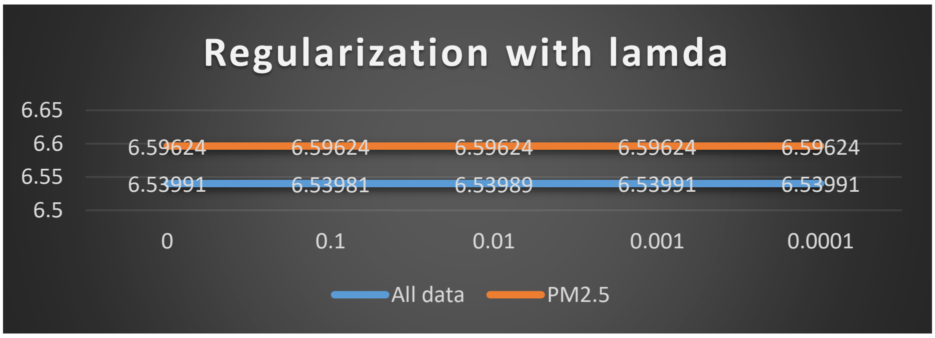
(1)抽所有汙染源 public:7.66887 private:5.41044 **ans:6.63643**

(2)抽 pm2.5 一次項 public:7.57904 private:5.79187 **ans:6.74491**

兩種模型準確率都比抽 9 個小時還要差，可能是因為 5 小時的 feature 數量較少，雖然在 training loss 可以達到 5.8 左右，但或許面臨者「Underfit」的問題，模型無法精確擬合 training data，因而影響到準確率

3. Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

由於實作的 loss function 只有一次方，因此 regularization 對我的實作並沒有太大的幫助，得出來的結果也與原本差不多，可能在小數點後面幾位才會發現差異。但若將 data 的維度次方增加，例如將 pm2.5 的 data 十次方，這樣的話 regularization 就會有避免 overfitting 的功用。



4. 在線性回歸問題中，假設有 **N** 筆訓練資料，每筆訓練資料的特徵 **(feature)** 為一向量 $\mathbf{x}^n$，其標註**(label)**為一存量 $\mathbf{y}^n$，模型參數為一向量 **w (**此處忽略偏權值 **b)**，則線性回歸的損失函數**(loss function)**為$\sum_{n=1}^{N}(\mathbf{y}^{n} - \mathbf{x}^{n} \cdot \mathbf{w})^2$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1\ \mathbf{x}^2 \ldots \mathbf{x}^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ \mathbf{y}^2 \ldots \mathbf{y}^N]^T$ 表示，請問如何以 **X** 和 **y** 表示可以最小化損失函數的向量 **w** ？請寫下算式並選出正確答案。**(**其中 $\mathbf{X}^T\mathbf{X}$ 為 **invertible)**

(c) $(X^TX)^{-1}X^Ty$

We want to find $\min \sum_{n=1}^{N} (y^n - x^n \cdot w)^2$

$\sum_{n=1}^{N} (y^n - x^n w)^2 = (Y - Xw)^T (Y - X\theta) = (Xw - Y)^T (X\theta - Y)$

Let cost function $J(w)$:

$J(w) = \frac{1}{2} \sum_{n=1}^{N} (y^n - x^n w)^2 = \frac{1}{2} [Xw - Y]^T [Xw - Y]$

∵ find the w to minimize $J(w)$

∴ find the w to let $\frac{\partial J(w)}{\partial w} = 0$

$\frac{\partial}{\partial w} J(w) = \frac{1}{2} \frac{\partial}{\partial w} (Xw - Y)^T (Xw - Y)$

$= \frac{1}{2} \frac{\partial}{\partial w} (w^T X^T X w - w^T X^T Y - Y^T X w + Y^T Y)$

$= X^T X w - X^T Y$

∵ $\frac{\partial J(w)}{\partial w} = 0$

∴ $X^T X w - X^T Y = 0$

$w = (X^T X)^{-1} X^T Y$ #