學號：B04901155 系級： 電機三 姓名：鄭閎

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. (2%)記錄誤差值 (RMSE)(根據kaggle public+private分數)，討論兩種feature的影響

	public	private	public+private
所有污染源	7.47992	5.30073	12.78065
pm2.5	7.46915	5.71105	13.1802

取所有污染源的結果比只取pm2.5的表現要好。我有把不同的污染源和pm2.5之間做consine similarity，發現其實都還滿高的，全部都在7.5以上，代表這些污染源的變動和pm2.5的相關性高。所以我想取所有污染源來訓練並預測pm2.5的數值是合理的。

此外觀察用所有污染源訓練出來的參數可以發現常有其他污染源的 weight 比pm2.5還要高的情形，下圖是取自9個小時的最後一小時的 weight。CO的weight就比pm2.5還高，在前幾個小時的結果也不乏pm2.5低於其他污染源的情形。只用pm2.5來估測讓模型包含的函數變少、彈性降低，而少包含到的函數有更好的結果。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

	public	private	public+private
所有污染源	7.68667	5.32697	13.01364
pm2.5	7.59362	5.84282	13.43644

整體表現抽前5小時都比抽前9小時表現差，觀察兩種情形訓練出的參數。

首先比較所有污染源的結果，為了比較多抽的前4的小時有什麼影響，我比較兩種方法抽的最早那個小時的參數（抽9小時是第1小時，抽五小時是第5小時）。

所有污染源		pm2.5	
0.0159283031 AMB_TEMP -0.0406109824 CH4 0.1338280105 CO -0.2891787789 NMHC 0.0301484461 NO 0.0229557544 NO2 0.0058414683 NOx 0.0027104653 O3 0.0071144627 PM10 -0.0377028041 PM2.5 0.0326269431 RAINFALL -0.0072325116 RH -0.2698701813 SO2 -0.0150967267 THC -0.0004689205 WD_HR -0.0019769969 WIND_DIREC -0.1348465522 WIND_SPEED -0.028329312 WS_HR	-0.0483366942 AMB_TEMP 0.0855094462 CH4 0.1952867086 CO 0.3668693352 NMHC 0.0456909982 NO -0.0511329149 NO2 -0.0001200112 NOx -0.0154199492 O3 0.002178577 PM10 -0.0529622982 PM2.5 -0.0470227304 RAINFALL -0.0330086881 RH -0.0397963793 SO2 0.1170849384 THC 0.0004859088 WD_HR 0.0006309452 WIND_DIREC -0.0583412226 WIND_SPEED -0.1082120829 WS_HR	-0.0345938251 -0.0176494651 0.2187903723 -0.2331413668 -0.0597500132 0.509722336 -0.5564234444 0.028821675 1.0700231628	-0.0937328112 0.40943424 -0.4605177922 0.0072328017 1.0599790969
抽9小時的第1小時	抽5小時的第1小時	第1~9小時的w	第1~5小時的w

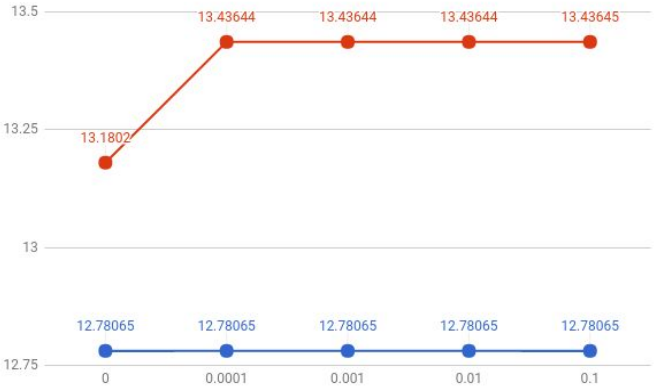
可以看到左圖中即使是離最後要預測的第10小時很遠，但是訓練出的weight和右圖比較還是有一些差不多甚至更大的影響力。所以只抽前5個小時丟棄的前4小時的資料造成結果較差有可能和上題說的一樣，損失了一些可能比較好的函數。

接著比較pm2.5的訓練結果。

發現第3、4天的參數比第5、8天的參數還更重要，因此沒有抽到這兩天來訓練真的忽略了重要的影響因子導致結果較差。

3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

	lambda	0	0.1	0.01	0.001	0.0001
all pollutants	public	5.30073	5.30073	5.30073	5.30073	5.30073
	private	7.47992	7.47992	7.47992	7.47992	7.47992
	sum	12.78065	12.78065	12.78065	12.78065	12.78065
pm2.5	public	5.71105	5.84282	5.84282	5.84282	5.84282
	private	7.46915	7.59363	7.59362	7.59362	7.59362
	sum	13.1802	13.43645	13.43644	13.43644	13.43644



圖中藍色為所有污染源、紅色為pm2.5。結果不同的regularization並沒有太大的差異。推測原因是模型只取一次項，函數完全是線性的，regularization讓函數更平滑的效果並不顯著。

4. (1%)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 x^n，其標註(label)為一純量 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 $\sum_{n=1}^{N}(y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2\ \dots\ x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2\ \dots\ y^N]^T$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。

(a) $(X^TX)X^Ty$

(b) $(X^TX)^{-0}X^Ty$

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

(d) $(X^TX)^{-2}X^Ty$

求 gradient descent 降到最低點的w，gradient為0

$$\frac{\partial L(w)}{\partial w} = -X^T(y - Xw) = 0$$

分配律

$$-X^Ty + X^TXw = 0$$

$$w = (X^TX)^{-1}X^Ty$$

答案選(c)