學號：R06942048 系級： 電信碩一 姓名：林彥伯

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	方均根
All feature	7.83	5.50	6.76605129
Only pm2.5	7.44	5.62	6.59310246

可以發現到只取 pm2.5 的效果會稍微好一點，因為全取 18 種的 feature 可能會取到一些沒有意義的 feature，導致在訓練時加入沒有必要的 weight 會使效果較差，在後續實驗我有試著只取某些較為重要的 feature eg: pm2.5 、pm10 、O3 效果又會比只取 pm2.5 小一點

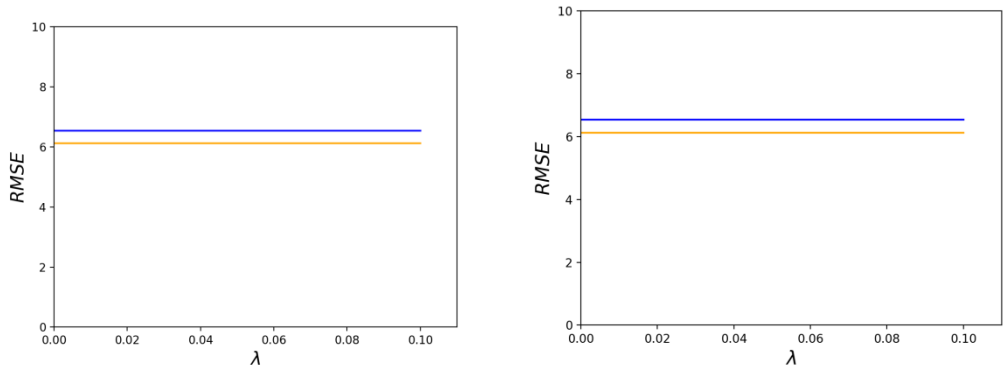
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	Public	private	方均根
All feature	7.737	5.378	6.66273416
Only pm2.5	7.579	5.791	6.7445134

只取五個小時可以發現到，error 沒有很明顯上升，甚至有一些下降，可以推測出不需要用到九個小時這麼多資料，也可以有不錯的預測效果，可以推測出下個時間點的 pm2.5 值只會與前幾個小時有關，不會到九小時那麼多

2. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

藍色線是 kaggle 上 test rmse 橘色線是 training



Only pm2.5　　all feature

由此圖可以推斷，λ regularization 在此次的設定中並沒有太大的效果，我有試著把 λ 調成 1 、10 但效果並沒有比較好，error 甚至上升

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)為一存量 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 n=1Nyn-xnw2 。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。(其中 X^TX 為 invertible)

a. $(X^TX)X^Ty$
b. $(X^TX)^{-0}X^Ty$
c. $(X^TX)^{-1}X^Ty$
d. $(X^TX)^{-2}X^Ty$

$$L(\vec{w}) = \sum_{n=1}^{N} (y^n - \vec{x}^n \vec{w})^2$$

$$\frac{dL}{dw} = 0 = -2\sum_{n=1}^{N} (y^n - \vec{x}^n \vec{w})\vec{x}^n$$

$$\Rightarrow x^T y - x^T x \vec{w} = 0$$

$$\Rightarrow x^T x \vec{w} = x^T y$$

$$\Rightarrow \vec{w} = (x^T x)^{-1} x^T y$$