學號：B07901069 系級： 電機一 姓名：劉奇聖

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 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

c. 第 1-3 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. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

答：

	全部汙染源		只有 pm2.5	
Learning rate	private	public	private	public
10	7.42838	6.04009	7.22356	5.90263
1	7.42658	6.03786	7.22356	5.90263
0.1	7.41223	6.01854	7.22357	5.90263

上表是使用 adagrad 依據兩種 feature 配合不同的 learning rate，iteration 10000 次，在 kaggle 上的誤差值。

從表中可看出只有 pm2.5 feature 的模型無論在 private 或 public 上都勝過全部汙染源的結果，推測可能原因是只有 pm2.5 feature 的模型參數較少，比較容易訓練到收斂(從上表可看出用 10~0.1 的 learning rate 訓練出來的模型分數幾乎相同)，也有可能是因為有全部汙染源的模型在訓練時需要用到更多的參數，導致其較只有 pm2.5 feature 的模型更容易 overfit 在 training data 上，使它在 testing 時的表現不佳。

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

答：

	全部汙染源		只有 pm2.5	
Learning rate	private	public	private	public
10	7.21188	6.06488	7.22552	6.22732
1	7.21054	6.06374	7.22552	6.22732
0.1	7.20084	6.05491	7.22552	6.22732

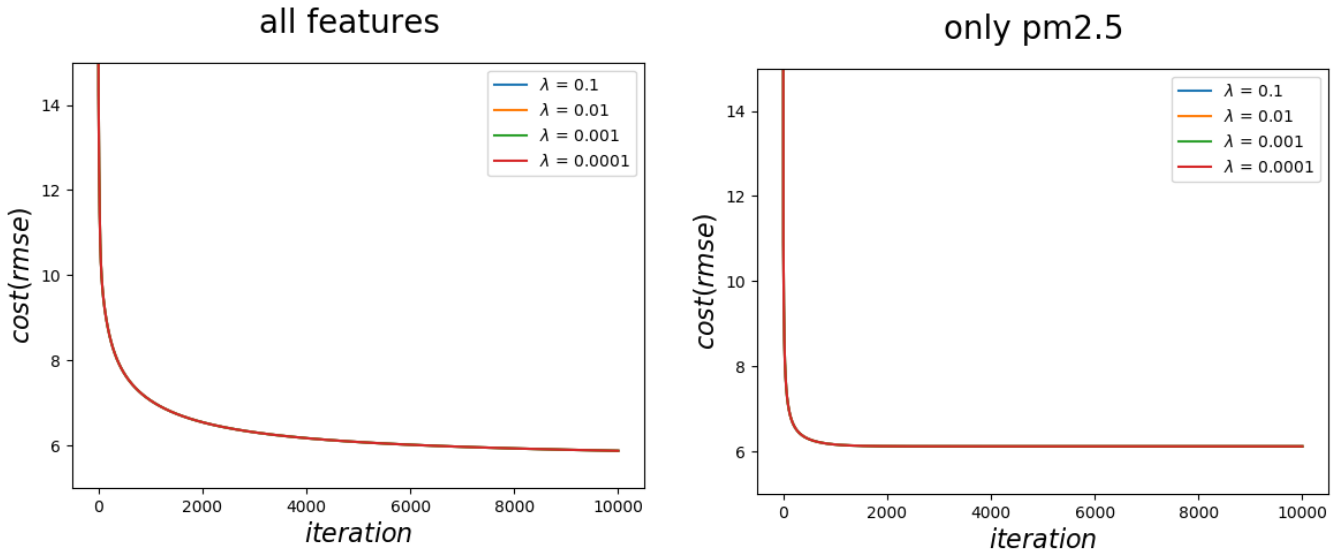
上表和第一題一樣是使用 adagrad 依據兩種 feature 配合不同的 learning rate，iteration 10000 次，在 kaggle 上的誤差值，唯一差別是改抽前 5 小時的 feature。

比較此表格與上題的表格可發現抽全部汙染源的模型在 private 上有進步，public 則無太明顯的差別，但整體(private+public)為進步，推測可能原因是原本抽 9 小時的模

型參數太多(18x9+1=163)，導致其容易對 training data overfit，改抽 5 小時後，參數減少(18x5+1=91)，降低了 overfit 的機率，因而在 testing 時有進步。

在只抽 pm2.5 的模型方面，雖然抽 5 小時和抽 9 小時兩種模型在 private 上的表現差不多，但抽 5 小時的模型在 public 上的表現卻較 9 小時的退步，因此整體(private+public)的表現為退步，推測可能原因是原本抽 9 小時的模型至少還有 9+1=10 的參數，但抽 5 小時的模型只有 5+1=6 個參數，模型太過簡單以至於發生 underfit(訓練時抽 5 小時確實比抽 9 小時在 training set 上的正確率低)，導致測試結果退步。

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



上面兩張圖是使用 adagrad 依據兩種 feature 配合不同的 λ 所做出來的 cost-iteration 圖，iteration 為 10000 次，抽的 feature 皆是連續 9 小時。

由於誤差太近，因此圖上看不出明顯的分離，但根據訓練時的誤差發現 λ 若越大則在 training set 上的 rmse 會稍微較高。L2 norm 是一種 weight decay，雖然會使訓練出來的模型較不容易 overfit，但相對來說就好像是減少了 iteration 的次數一樣，因此在 training set 上的 rmse 會較高。

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 ？請選出正確答案。(其中 X^TX 為 invertible)

(a) $(X^TX)X^Ty$
(b) $(X^TX)yX^T$
(c) $(X^TX)^{-1}X^Ty$
(d) $(X^TX)^{-1}yX^T$

答：(c)