學號：B05901003 系級： 電機二 姓名：徐敏倩

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

	Kaggle Private Score	Kaggle Public Score	Average RMSE
All features used	7.19949	5.64897	6.39404020564463
Only pm2.5 used	7.23586	5.92746	6.61409298359193

由上表可知，只有抽取pm2.5的一次項當作feature時的誤差較抽取全部污染源作feature時要差，然而相差並不大。因此，可推知pm2.5提供了大部分的資訊，但其他的污染源對於pm2.5的數值依然有些影響，故誤差依然較小。

(learning rate 為1、gradient descent 的部分用adagrad、iterations為200000)

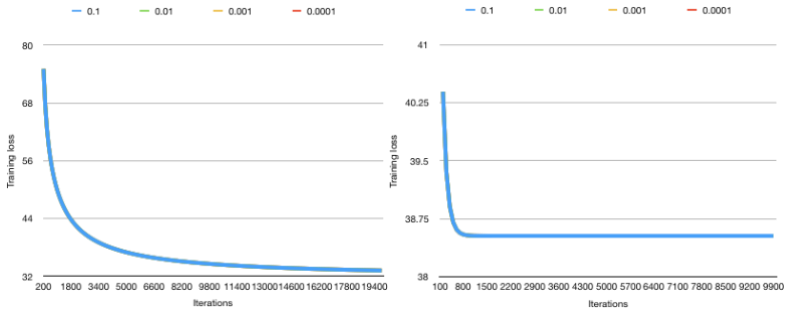
2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

	Kaggle Private Score	Kaggle Public Score	Average RMSE
All features used: 9 hours	7.19949	5.64897	6.47083913882118
Only pm2.5 used: 9 hours	7.23113	5.90928	6.60336396071351
All features used: 5 hours	7.17426	5.95704	6.59379754425324
Only pm2.5 used: 5 hours	7.25806	6.20458	6.75189780506192

由上表可知，取前5小時相較於取前9小時，不論是在抽取所有污染源抑或是只取pm2.5作feature的模型中，表現都比較差。由此可知，只抽取前5小時的feature會刪掉一些對於預測結果有用的資料，使得計算出來的誤差較原本的為大。

(learning rate 為1、gradient descent 的部分用adagrad、iterations為200000)

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



All features used　　Only pm2.5 used

由上圖可知，不論是用所有的污染源當feature，抑或是單用pm2.5作feature，λ得值對於不同iteration時的training loss，沒有什麼影響，不論λ為何，在同一個iteration的training loss都有幾近相同的值，相差都只在10^{-6}左右。可知對於此training data而言，regularization term 對於training loss並無什麼影響。

(learning rate 為1、gradient descent 的部分用adagrad)

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

Ans. (c)

Loss function = $\sum_{n=1}^{N}(y^n - x^n \cdot w)^2$

$$\frac{\partial \text{Loss function}}{\partial \mathbf{w}} = 2\sum_{n=1}^{N}(y^n - x^n \cdot w)\cdot(-x^n) = -2X^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

當 w 為最小化損失函數的向量時，$\partial \text{Loss function}/\partial \mathbf{w}$為0

故，

$2X^TX\mathbf{w} = 2X^T\mathbf{y}$，且$X^T\mathbf{X}$為invertible

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