

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  $p = 9 \times 18 + 1$  而(2) 代表  $p = 9 \times 1 + 1$

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

答:

	Public score	Private Score
所有 features	5.63779	7.21546
只有 PM2.5	5.90263	7.22356

由實驗結果可知，只有 PM2.5 的預測誤差值較所有 features 誤差值大。這有可能是影響當前的 PM2.5 數值不僅只有先前的 PM2.5 數值，也包含其他元素的影響。由維基百科的資料顯示，PM2.5 的成分可能由硫和氮的氧化物轉化而成，因此空氣中硫和氮的氧化物濃度可能與 PM2.5 數值有關。

(備註: 所有實驗皆使用 Adagrad; 初始學習率 1.0; Iterations: 10000)

(資料來源: <https://zh.wikipedia.org/wiki/%E6%87%B8%E6%B5%AE%E7%B2%92%E5%AD%90>)

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

答:

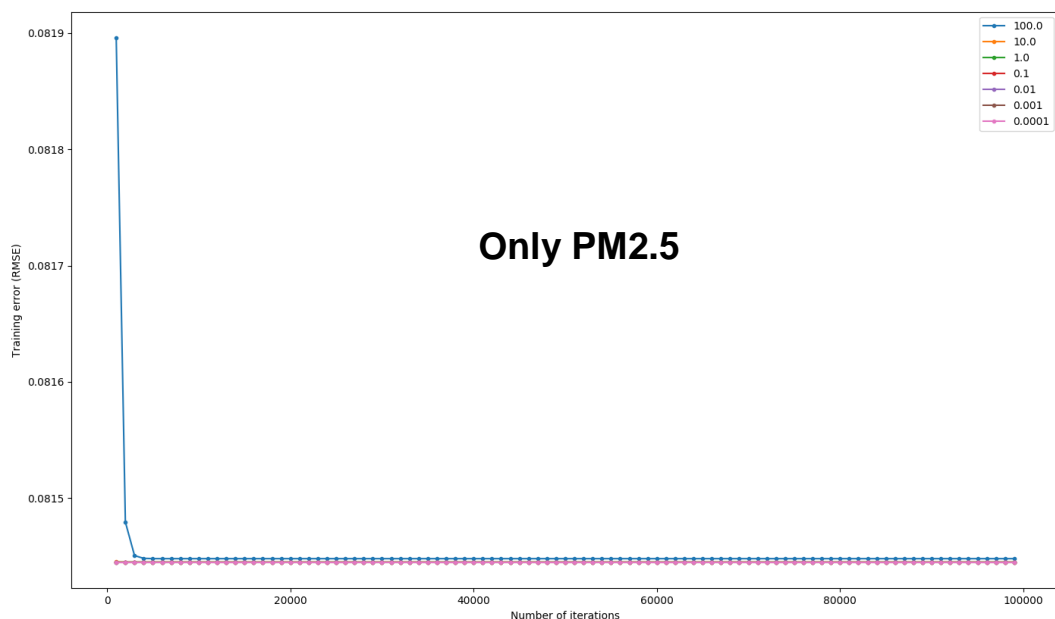
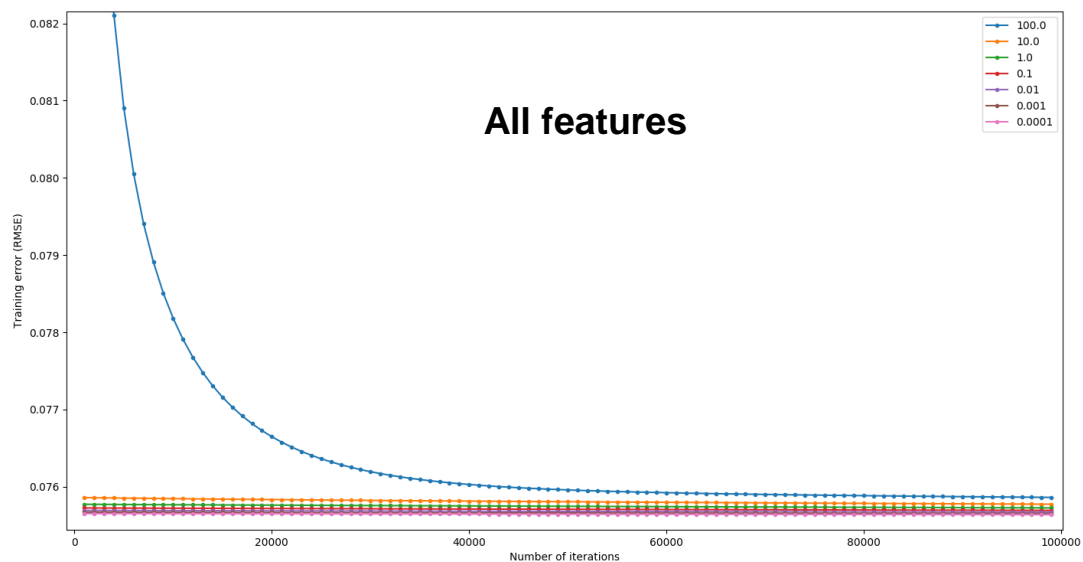
	Public score	Private Score
所有 features + 抽前 9 小時	5.63779	7.21546
所有 features + 抽前 5 小時	5.98257	7.16701
只有 PM2.5 + 抽前 9 小時	5.90263	7.22356
只有 PM2.5 + 抽前 5 小時	6.22732	7.22552

大致而言，抽取前 9 小時的誤差在 public 與 private set 上的誤差都比抽取前 5 小時的誤差小 (除了所有 features 在 private set 上的表現)。以這結果而言，當前 PM2.5 的值與先前的 PM2.5 有一定的相關性，因此參考較多小時前的值，有助於降低預測誤差值。另外，在訓練過程中，由於抽取前 5 小時的 features 較少，因此 training 的時間有顯著的降低。

(備註: 所有實驗皆使用 Adagrad; 初始學習率 1.0; Iterations: 10000)

### 3. (1%)Regularization on all the weights with $\lambda=0.1$ 、 $0.01$ 、 $0.001$ 、 $0.0001$ ，並作圖

答:



答: 由上二圖可知，當  $\lambda$  越大時，training error 越大，隨著  $\lambda$  漸小，training error 也隨之減小。我們知道 regularization 能使 model 對於 input 的變化減小，從而避免 overfitting 的狀況。因此當  $\lambda$  大時，model 並不會趨向符合 training data，這也就是為何  $\lambda$  越大，training error 越大。另外，在 all features 的實驗中，不同的  $\lambda$  對應的 error 有顯著的不同；但在 only PM2.5 的實驗中，不同的  $\lambda$  對應的 error 並沒有顯著的不同。這可能是因為 only PM2.5 的 model 已經非常簡單，不像 all features 的 model 那樣複雜，所以不同的  $\lambda$  對於 only PM2.5 的實驗並沒有太大的影響。

(備註: 所有實驗皆使用 Adagrad; 初始學習率 1.0; Iterations: 10000)

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  $X^T X$  為 invertible)

- (a)  $(X^T X) X^T y$
- (b)  $(X^T X) y X^T$
- (c)  $(X^T X)^{-1} X^T y$
- (d)  $(X^T X)^{-1} y X^T$

答: (c)