

學號：B05902128 系級：資工三 姓名：鄭百凱

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  $p = 9 * 1 + 1$

註：1~3 題皆是用 adagram, 其中 learning rate = 1，起始值都是 0, iterate 100000 次

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

	(1) $p = 9 \times 18 + 1$	(2) $p = 9 * 1 + 1$
Public	5.63779	5.90263
Private	7.21546	7.22356
平均	6.42663	6.56310

由表可見不管在 public 還是 private 的 data set，163 維的 linear regression 皆有著較低的 error。原因估計是因為 163 維相較 10 維有著較大的 function space，因此 bias 比較小。

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

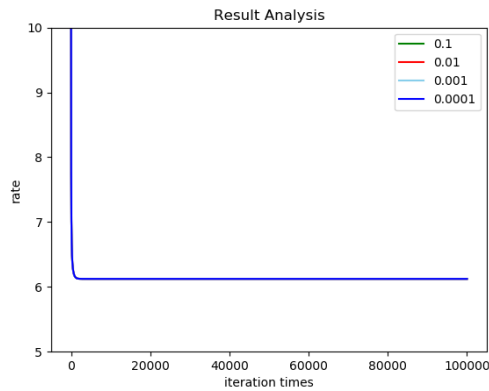
	(1) 5HR $p = 5 \times 18 + 1$	(2) 5HR $p = 5 * 1 + 1$
Public	5.98257	6.22732
Private	7.16701	7.22552
平均	6.57479	6.72642

抽前 5 小時的 error 比抽前 9 小時的 error 大，而且 5 小時中取全部的 feature 又比只取 PM2.5 要來的好。1、2 小題中，可以看出維度的數量大致跟 error 呈負相關。

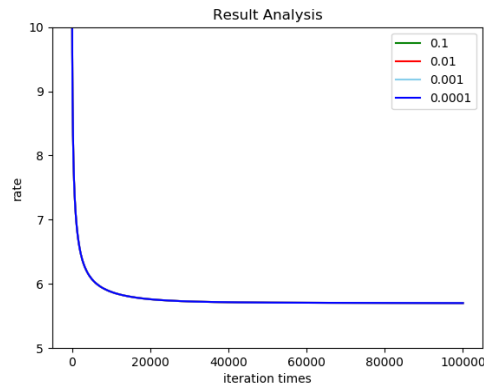
3. (1%)Regularization on all the weight with  $\lambda=0.1$ 、 $0.01$ 、 $0.001$ 、 $0.0001$ ，並作圖

將不同的 lambda 產生的 w 丟到 kaggle 的結果

	$\lambda=0.1$	$\lambda=0.01$	$\lambda=0.001$	$\lambda=0.0001$
Public (1) $p = 9 \times 18 + 1$	5.63780	5.63779	5.63779	5.63779
Private (1) $p = 9 \times 18 + 1$	7.21545	7.21546	7.21546	7.21546
平均 (1) $p = 9 \times 18 + 1$	6.42663	6.42663	6.42663	6.42663
Public (2) $p = 9 * 1 + 1$	5.90263	5.90263	5.90263	5.90263
Private (2) $p = 9 * 1 + 1$	7.22356	7.22356	7.22356	7.22356
平均 (2) $p = 9 * 1 + 1$	6.56310	6.56310	6.56310	6.56310



10 維



163 維

由圖可知，不管是 10 維還是 163 維，四條線都重疊了。所以在只改變  $\lambda$  且  $\lambda$  介於 0.1~0.0001 的情況下，error 的下降速率是幾乎一樣的。要看出變化可能要讓  $\lambda$  提高到 10 以上。而從圖或從表都可以看出來，就算加了 regularization，163 維的表現還是比 10 維好。

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

C

將 loss function 對  $w$  做微分，使其等於一個 0 向量。解完這個方程式就可以得到  $w$  的最佳解為  $(X^T X)^{-1}X^T y$