

Machine Learning HW1 Report

學號：b06901007 系級：電機二 姓名：戴子宜

2019.3.8

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. 紀錄誤差值(RMSE) (根據 kaggle public+private 的分數)，討論兩種 feature 的影響。

model (1)：抽全部 9 小時內的污染源 feature 當作一次項（加 bias）

model (2)：抽全部 9 小時 pm2.5 的一次項當作 feature（加 bias）

	Training set	Kaggle (Public)	Kaggle (Private)	Kaggle (Average)
model (1)	5.78290	5.81765	7.28173	6.59047
model (2)	6.20711	5.93022	7.24763	6.62176

由結果可以看出來只抽取 pm2.5 的 model 在 training set 或是 testing set 的 loss 都比較大。在 training set 的 loss model (2) 較 model (1) 大蠻多的是因為 model (2) 的參數比較少，所以整個 function set 比較小。在 testing set 的 loss 兩個 model 其實差距很近，但 model (1) 表現略好一點，代表 model (1) 雖然參數較多，training loss 較少，但沒有因為考慮較多參數而 overfit，代表除了 pm2.5 以外的某些參數也是有助於預測 pm2.5 的。

Average 的算法為 $\sqrt{(Public^2 + Private^2)/2}$ ，是整個 test set 的 RMSE。

2. 將 feature 從抽前 9 小時改就抽前 5 小時，討論其變化。

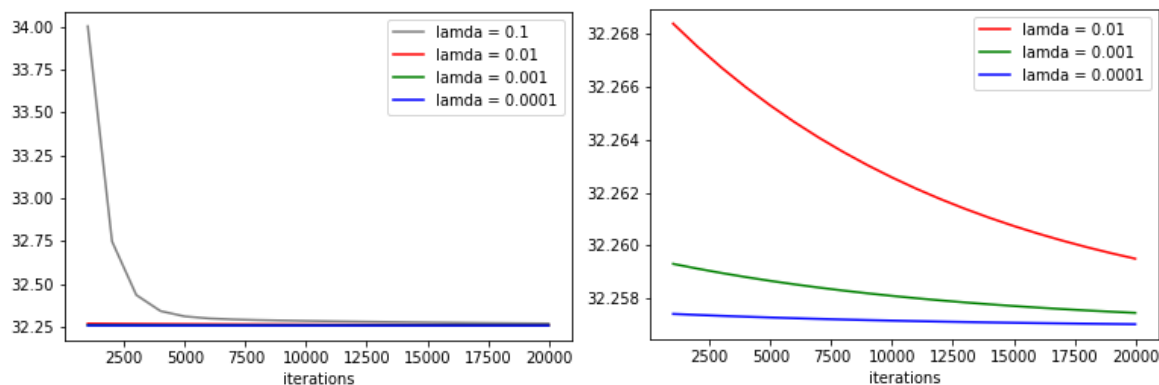
model (3)：抽前 5 小時內全部的污染源 feature 當作一次項（加 bias）

model (4)：抽前 5 小時內 pm2.5 的 feature 當作一次項（加 bias）

	Training set	Kaggle (Public)	Kaggle (Private)	Kaggle (Average)
model (1)	5.78290418	5.81765	7.28173	6.59047
model (2)	6.20711447	5.93022	7.24763	6.62176
model (3)	5.90772849	6.01045	7.24285	6.65523
model (4)	6.29138864	6.23692	7.24512	6.75984

由結果可以看出來抽取前 5 小時的 model 在 training set 或是 testing set 的誤差都比較大，代表抽取前 9 小時的 model 雖然參數較多，training loss 較少，但沒有因為考慮較多參數而 overfit，因此除了第 6~9 天前的污染源也有助於預測 pm2.5。

3. Regularization on all weight with $\lambda=0.1$ 、 0.01 、 0.001 、 0.0001 ，並作圖。



圖的縱軸為 training loss，從圖可以發現當 λ 越大，需要花比較多 iteration 才會 train 到收斂，且當 λ 越大，最後收斂時的 train loss 也會比較大，這是因為加入 regularization 可以避免 overfit。

（上圖的 training model 為 model (1)，learning rate 的初始值為 1，gradient descent 的部分是利用 adagrad）

4. 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)

Ans: (c) $(X^T X)^{-1} X^T y$