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 = 9x18+1 而(2) 代表 p = 9*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）

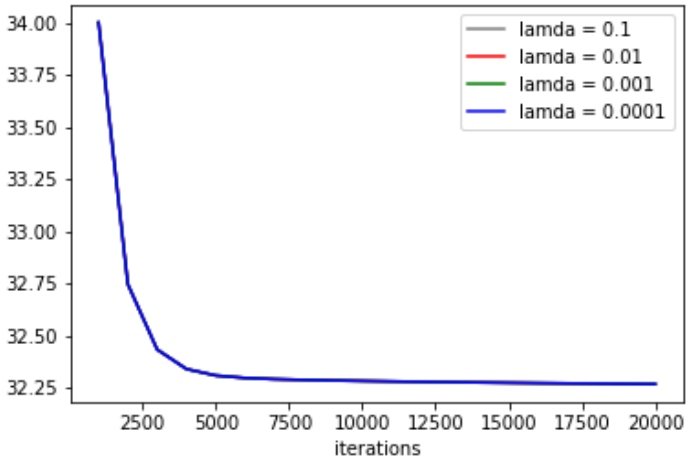
	Training set	Kaggle (Public)	Kaggle (Private)	Kaggle (Average)
model (1)	5.78290418	5.81765	7.28173	6.59047
model (2)	6.20711447	5.93022	7.24763	6.62176
model (3)	5.90772849	6.01045	7.24285	6.65523
model (4)	6.29138864	6.23692	7.24512	6.75984

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 λ=0.1、0.01、0.001、0.0001，並作圖。

iterations	lamda = 0.1	lamda = 0.01	lamda = 0.001	lamda = 0.0001
999	34.001978721402224	34.00119856946806	34.001120584668975	34.00111278649308
1999	32.74769841493631	32.746894787871284	32.746814493599146	32.746806464856654
2999	32.43515812543827	32.43451055866816	32.434445930058665	32.434439468479574
3999	32.3414555985215	32.34099662580839	32.340950923435585	32.34094635514968
4999	32.31036584958577	32.31005804206037	32.310027516364606	32.310024466349404
5999	32.29844808511348	32.29824238781538	32.29822212074257	32.29822009706706
6999	32.292711374020264	32.2925681227093	32.29255413502569	32.29255273963823
7999	32.289117834619596	32.28900983671421	32.28899939851346	32.288998358316725
8999	32.28636832146521	32.28627835947606	32.286269741020156	32.286268882960265
9999	32.284026126084775	32.283943776792476	32.2839359301179	32.28393514934188
10999	32.28193409201036	32.281853235117396	32.28184554437177	32.281844779256005
11999	32.28002881975396	32.27994596365093	32.2799380771709	32.27993729252395
12999	32.278279799243705	32.27819303086641	32.2781847557685	32.27818393228613
13999	32.27666883398803	32.27657718208319	32.27656842030204	32.27656754816829
14999	32.27518282102645	32.275085876090984	32.2750765861719	32.275075661236215
15999	32.27381109722503	32.273708785723926	32.273698960107026	32.27369798161135
16999	32.27254440262831	32.272436852491026	32.27242650396812	32.272425473191625
17999	32.27137443982477	32.27126189962353	32.271251053193325	32.27124997263731
18999	32.27029366340937	32.2701764531267	32.27016514102914	32.27016401392004
19999	32.269295163364944	32.26917364385583	32.269161902489216	32.26916073247002

表格中的數字代表 training MSE error。從圖可以發現當λ越大，在相同的 iteration 的 MSE loss 會稍大一點，但其實數據的差距不明顯，畫在圖上幾乎會重疊。



（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 $X = [x^1\ x^2 \ldots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \ldots y^N]^T$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) $(X^T X)^{-1} X^T y$