Machine Learning 2019 Spring - HW1 Report

學號：B06902049　　系級：資工二　　姓名：林首志

（感謝裴梧鈞（B06902029）同學提供漂亮的Markdown格式Report模板）

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

Model	Public	Private	Average
Model 1	5.77421	7.27931	6.52676
Model 2	5.90263	7.22356	6.56310

從表格中的平均誤差可以看出Model 1預測的比Model 2還要精準一些。（由於測試資料量極少，在這裡不單獨討論Public或Private的表現）由於Model 2使用的features比Model 1少，資訊量相對少了一些，因此Model 2的平均誤差比Model 1高是合理的（有時候會有反例，例如加入一些不太相關的資料可能會使結果更差）。另一種觀點是，Model 2的Hypothesis set是Model 1的Hypothesis set的子集，也就是說Model 1的Hypothesis set比Model 2還要powerful，因此Model 1的確有可能（在沒有過擬合的情況下）做出比Model 2更好的結果。

2. (1%) 將feature從抽前9小時改成抽前5小時，討論其變化

Model	Public	Private	Average
Model 1	5.99388	7.23626	6.61507
Model 2	6.22749	7.22464	6.72607

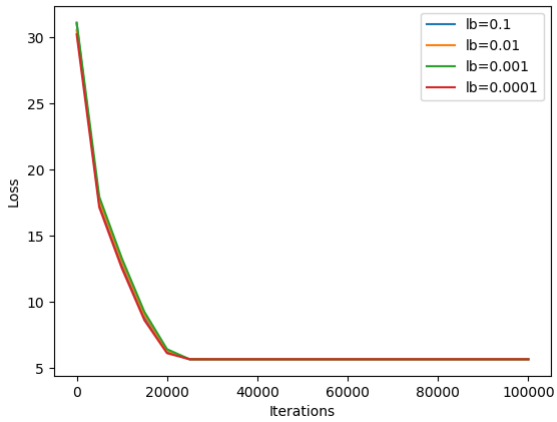
由於使用的資料量更少，兩個Model的預測精確度都下降了，但是Model 1的表現仍然比Model 2好。

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

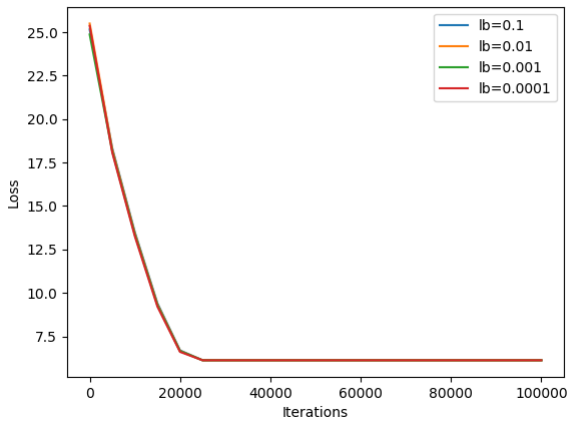
（在此沒有對bias做regularization，圖中Loss(RMSE)不包含regularization term，training使用Adam）

以下是iteration - training loss的折線圖：

Model1：



Model2：



從圖可以推論，對於這樣的Model、training data、training algorithm及這些lambda取值，training過程的loss是相近的。

在測試資料的表現(Public和Private的平均)如下表：

Model	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$	$\lambda = 0.0001$
Model 1	6.52563	6.52665	6.52676	6.52676
Model 2	6.56355	6.56314	6.56310	6.56310

從表可以看出Model 1的表現隨著λ增加而變好，而Model 2的表現則相反。可能的原因是Model 1原本處在overfitting的狀態，regularization會讓表現更好，而Model 2原本處在underfitting的狀態，regularization只會讓表現更差。

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 $\mathbf{w}$ (此處忽略偏權值 $\mathbf{b}$)，則線性回歸的損失函數(loss function)為 $\sum_{n=1}^{N}(\mathbf{y}^n - \mathbf{x}^n\mathbf{w})^2$。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1\ \mathbf{x}^2\ \dots\ \mathbf{x}^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ \mathbf{y}^2\ \dots\ \mathbf{y}^N]^T$ 表示，請問如何以 $\mathbf{X}$ 和 $\mathbf{y}$ 表示可以最小化損失函數的向量 $\mathbf{w}$？請選出正確答案。(其中 $\mathbf{X}^T\mathbf{X}$ 為invertible)

    1. $(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}$
    2. $(\mathbf{X}^T\mathbf{X})\mathbf{y}\mathbf{X}^T$

3. $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
4. $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{y}\mathbf{X}^T$

Answer : 3

基本的概念是在題目給定的條件下($\mathbf{X}^T\mathbf{X}$可逆)，其Loss function是凸函數且有唯一的最低點。而在最低點的必要條件是梯度等於零，經過一番推導可以得出最好的$\mathbf{w}$滿足$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$，將$\mathbf{X}^T\mathbf{X}$移項即可得到最佳解的公式。