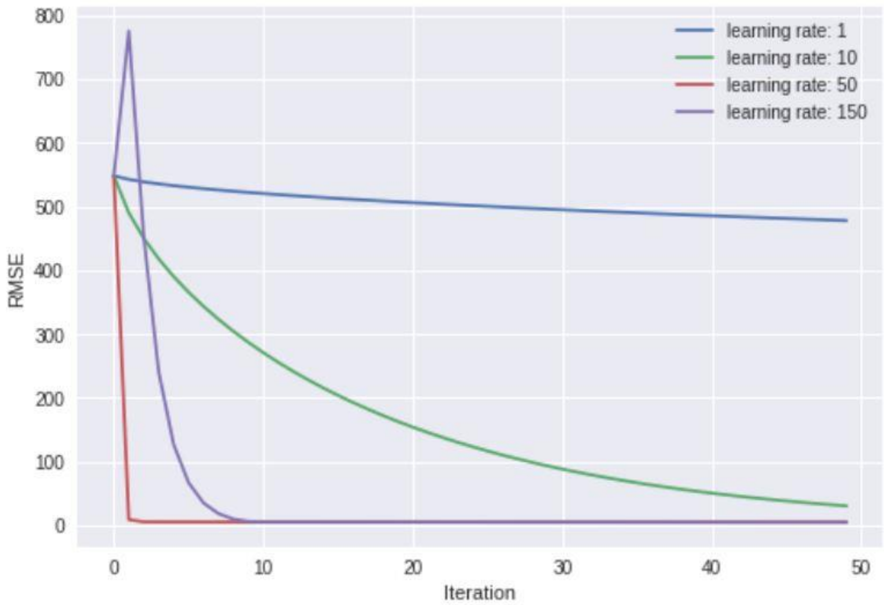
Homework 1 Report - PM2.5 Prediction

學號： B05602022 系級：工海三 姓名: 盧庭偉

1. (1%) 請分別使用至少 4 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。

以下為示意圖：



圖中可看見，當 learning rate 太小時，需要走很長的時間。而 learning rate 太大則有可能會出現震盪現象。由於本次範例有使用 Adagrad，後面走的步伐會變小，因此 learning rate 太小有可能導致走不到 minimum 的位置，而 learning rate 太大則不太會有影響因為當步伐越來越小後，它會慢慢找到適當的位置(有點類似顯微鏡的概念。先用粗調節輪找觀察物，再用細調節輪微調焦距)。因此在有使用 Adagrad 的情況下，learning rate 寧願太大也不要太小。

2. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

	Regularization	Kaggle public score
9 小時內 PM2.5	$\lambda = 0$	9.22144
9 小時內 PM2.5	$\lambda = 50$	9.24584

只使用 PM2.5 當資料的情況下，加入 regularization 並沒有使結果更好，判斷應為 underfitting。

9 小時內所有 feature	$\lambda = 0$	9.02981
9 小時內所有 feature	$\lambda = 50$	8.74607

使用全部的 feature 下去 train 的情況下，分數有所進步，證實了前面的假設(只使用 PM2.5 會 underfit)。加入 regularization 能使結果略微進步，判斷可能有些許 overfit。

3. (1%)請分別使用至少四種不同數值的 regularization parameter λ 進行 training（其他參數需一致），討論及討論其 RMSE(training, testing)（testing 根據 Kaggle 上的 public/private score）以及參數 weight 的 L2 norm。

Regularization	RMSE(Training)	Public Score	Private Score	L2 norm
$\lambda = 1$	3.96	8.36351	6.90362	1.17
$\lambda = 50$	3.96	8.35447	6.90897	1.08
$\lambda = 100$	3.96	8.34457	6.91661	0.97
$\lambda = 200$	3.96	8.33487	6.94574	0.45

由於 training data 中存在許多有問題的資料，連續的 0 (空資料)、負的參數及不可能的數據 (PM2.5 > 900)，因此在 training 前我已把這些資料去除。方法是先刪掉連續為零以及負的資料，剩下的再去除三個標準差之外的資料。

但也因為這個事前處理，我們可能會把一些極端但合理的資料去除，因此我們的 loss function 應該會比較平滑。而測試了幾個 regularization parameter 之後，結果並沒有顯著的進步，也算是驗證了我們的假設。

Problem 4~6 collaborator:

R06525062 吳政道
TA

4.

(4-a)

定義 R 為一 $N \times N$ 矩陣，對角線為 $r_1, r_2, \cdots, r_N$，其餘為 0

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} r_n (t_n - w^T x_n)^2$$

$$= \frac{1}{2}(t_n - w^T x_n) R (t_n^T - x_n^T w)$$

$$= \frac{1}{2}(t_n R t_n^T - t_n R x_n^T w - w^T x_n R t_n^T + w^T x_n R x_n^T w$$

$$= \frac{1}{2}(t_n R t_n^T - 2 t_n R x_n^T w + w^T x_n R x_n^T w)$$

$$\frac{\partial}{\partial w} E_D(w^*) = x_n R x_n^T w^* - t_n R x_n^T = 0$$

$$\Rightarrow w^* = (x_n R x_n^T)^{-1} t_n R x_n^T = (x_n R x_n^T)^{-1} x_n R t_n^T \quad \#$$

(4-b)

$$w^* = \left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \cdot \begin{bmatrix} 2 & & \\ & 1 & \\ & & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & & \\ & 1 & \\ & & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$= \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}^{-1} \begin{bmatrix} 125 \\ 100 \end{bmatrix} = \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix} \begin{bmatrix} 125 \\ 100 \end{bmatrix}$$

$$= \begin{bmatrix} 2.28 \\ -1.14 \end{bmatrix} \quad \#$$

5.

NO.

DATE / /

5. $$E'(w) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, w+\Delta w) - t_n\right)^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(w_0 + \sum_{i=1}^{D} w_i (x_n + \Delta x_i) - t_n\right)^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left[\left(w_0 + \sum_{i=1}^{D} w_i x_i - t_n\right)^2 + 2\left(w_0 + \sum_{i=1}^{D} w_i x_i - t_n\right)\left(\sum_{i=1}^{D} w_i \Delta x_i\right) + \left(\sum_{i=1}^{D} w_i \Delta x_i\right)^2\right]$$

$$\mathbb{E}(E'(w)) = \mathbb{E}(E(w)) + \frac{1}{2}\sum_{n=1}^{N}\left[2\left(w_0 + \sum_{i=1}^{D} w_i x_i - t_n\right)\left(\sum_{i=1}^{D} w_i \mathbb{E}(\varepsilon_{ni})\right) + \mathbb{E}\left(\sum_{i=1}^{D} w_i \varepsilon_i\right)^2\right]$$

$$\because \begin{cases} \mathbb{E}(\varepsilon_i) = 0 \\ \mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \varepsilon_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i \varepsilon_i w_j \varepsilon_j\right] = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij} \sigma^2 = \sigma^2 \sum_{i=1}^{D} w_i^2 \end{cases}$$

$$\therefore \mathbb{E}(E'(w)) = \mathbb{E}(E(w)) + \frac{1}{2}\sum_{n=1}^{N} \sigma^2 \sum_{i=1}^{D} w_i^2$$

$$= \mathbb{E}(E(w)) + \frac{N\sigma^2}{2}\sum_{i=1}^{D} w_i^2$$

$$= \mathbb{E}(E(w)) + \lambda \sum_{i=1}^{D} w_i^2 \text{, where } \lambda = \frac{N\sigma^2}{2} \quad \#$$

6.

DATE / /

6. $$\frac{d}{d\alpha}\ln|A| = \frac{d}{d\alpha}\ln(\lambda_1 \cdot \lambda_2 \cdots \lambda_N) \text{, where } \lambda_i \text{ are eigenvalues of } A$$

$$= \frac{d}{d\alpha}\sum_{i=1}^{N}\ln\lambda_i = \sum_{i=1}^{N}\frac{1}{\lambda_i}\frac{d\lambda_i}{d\alpha} \cdots ①$$

$$\because Au = \lambda u \Rightarrow A^{-1}u = \frac{1}{\lambda}u \Rightarrow A^{-1}\frac{dA}{d\alpha}u = \frac{1}{\lambda}\frac{d\lambda}{d\alpha}u$$

$$\therefore \operatorname{Tr}\left(A^{-1}\frac{d}{d\alpha}A\right) = \sum_{i=1}^{N}\frac{1}{\lambda_i}\frac{d\lambda_i}{d\alpha} = ① = \frac{d}{d\alpha}\ln|A| \quad \#$$