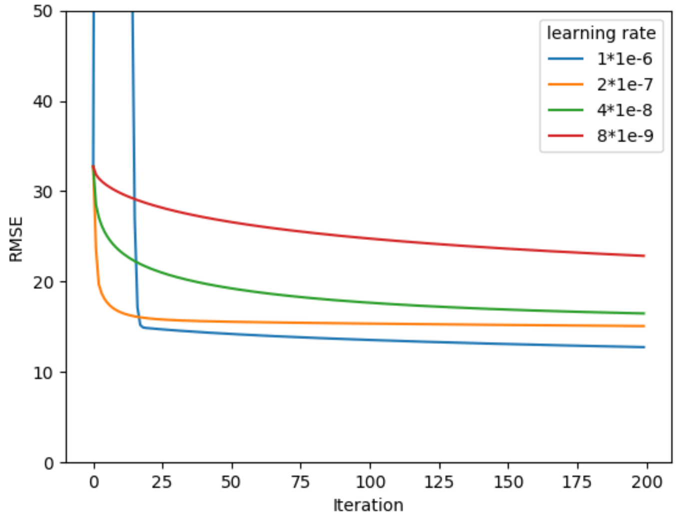
Homework 1 Report - PM2.5 Prediction

學號： r07922050 系級： 資工所碩一 姓名: 洪正皇

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。



當 learning rate 為 1e-6 的時候，前幾次 iteration 可以看到 RMSE 衝的很高，因為初始的 w 離最佳的 w*可能偏遠，導致 loss function 過大、使用 gradient descent 時 w 時更新過頭。而在修正回來的過程中，因為 learning rate 較大，在 25 次 iteration 之前就比其他三種 learning rate 更接近 w*。其餘的三種 learning rate，則是正常的接近最佳解 w*，越大的 learning rate 接近的越快。

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）。

使用所有 feature 的一次項：8.21580 / 8.65027 (public / private score)
使用 PM2.5 的一次項：6.59572 / 7.02786 (public / private score)
我認為有部分的 feature 與 PM2.5 相關性不大，而過少的資料量，導致不大相關的 feature 在 training 時得到的 weight 不夠小，最後在 testing 時影響了表現。此外，我對於 PM2.5 以外的資料，不確定如何做較佳的前處理，因此其他 feature 的雜訊也影響了 RMSE 的表現。

3. (1%)請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論及討論其 RMSE(traning, testing)（testing 根據 kaggle 上的 public/private score）以及參數 weight 的 L2 norm。

regulization parameter λ	RMSE(traning)	public/private score	L2 norm
300	16.424558	15.26717/ 15.00662	26.960319
100	10.686984	9.70320 / 10.05367	27.332276
10	10.107056	8.64379 / 9.30555	27.823405
1	10.107329	8.64317 / 9.30486	27.884081
0.1	10.107820	8.63708 / 9.29809	27.880857
0.01	10.107874	8.58879 / 9.24097	27.890718

lambda 在 10 以下時，在 training 時都有對 RMSE 有些微進步，但差距不大，因此在 test 時得到的成績也彼此差不多。另外 lambda 越大時，可以看到 L2 norm 有縮小，然而在 lambda 大於 100 之後，即使 L2 norm 縮小了，但 regularization 的影響讓 loss function 偏差太大，而不是向著最佳解前進，因此導致 RMSE 表現更差。

4~6 (3%) 請參考數學題目（連結：），將作答過程以各種形式（latex 尤佳）清楚地呈現在 pdf 檔中（手寫再拍照也可以，但請注意解析度）。

HackMD Link: https://hackmd.io/s/SJEjt369Q

下一頁

(4-a)

Let $X = [x_1 \;\; x_2 \;\; x_3 \;\; \dots \;\; x_n]$, $Y = [t_1 \;\; t_2 \;\; t_3 \;\; \dots \;\; t_n]^T$

and

$$R = \begin{pmatrix} r_1 & 0 & 0 & 0 & 0 \\ 0 & r_2 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & r_n \end{pmatrix}$$

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

$= \frac{1}{2}(Y^T - \mathbf{w}^T X)R(Y - X^T\mathbf{w})$

$= \frac{1}{2}(Y^T RY - \mathbf{w}^T XRY - Y^T RX^T\mathbf{w} + \mathbf{w}^T XRX^T\mathbf{w})$

Find $\mathbf{w}^*$ that minimizes the error function is equal to take $\mathbf{w}^*$ let $E_D(\mathbf{w}^* + \Delta\mathbf{w}) - E_D(\mathbf{w}^*) = 0$

$E_D(\mathbf{w}^* + \Delta\mathbf{w}) - E_D(\mathbf{w}^*)$

$= \frac{1}{2}(Y^T RY - (\mathbf{w}^* + \Delta\mathbf{w})^T XRY - Y^T RX^T(\mathbf{w}^* + \Delta\mathbf{w}) + (\mathbf{w}^* + \Delta\mathbf{w})^T XRX^T(\mathbf{w}^* + \Delta\mathbf{w}))$

$- \frac{1}{2}(Y^T RY - \quad \mathbf{w}^{*T} XRY - Y^T RX^T\mathbf{w}^* \quad + \mathbf{w}^{*T} \quad XRX^T\mathbf{w}^*)$

$= \frac{1}{2}(-\mathbf{\Delta w}^T XRY - Y^T RX^T\mathbf{\Delta w} + \Delta\mathbf{w}^T XRX^T\mathbf{w}^* + \mathbf{w}^{*T} XRX^T\Delta\mathbf{w} + \Delta\mathbf{w}^T XRX^T\Delta\mathbf{w})$

$= \frac{1}{2}(-2\mathbf{\Delta w}^T XRY + 2\Delta\mathbf{w}^T XRX^T\mathbf{w}^* + \Delta\mathbf{w}^T XRX^T\Delta\mathbf{w})$

$\because$

$=> \Delta\mathbf{w}^T(-XRY + XRX^T\mathbf{w}^* + \frac{1}{2}XRX^T\Delta\mathbf{w})$

$= 0$

$=> -XRY + XRX^T\mathbf{w}^* = 0$

$=> \mathbf{w}^* = (XRX^T)^{-1}XRY$

(4-b)

$\mathbf{w}^* =$

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

$$= \begin{bmatrix} \frac{5175}{2267} \\ \frac{-2575}{2267} \end{bmatrix}$$

5

Let $\tilde{x} = x + \epsilon$

The minimizing E averaged over the noise distribution :

$\mathbb{E}[E(\mathbf{w})]$
$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\big(y(\tilde{x}_n, \mathbf{w}) - t_n\big)^2]$
$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\big(y(x_n, \mathbf{w}) - t_n + \sum_{i=1}^{D} w_i\epsilon_i^{(n)}\big)^2]$
$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\big[\big(y(x_n, \mathbf{w}) - t_n\big)^2 + 2\sum_{i=1}^{D} w_i\epsilon_i^{(n)}\big(y(x_n, \mathbf{w}) - t_n\big) + \big(\sum_{i=1}^{D} w_i\epsilon_i^{(n)}\big)^2\big]]$
$= \frac{1}{2}\sum_{n=1}^{N}\big[\big(y(x_n, \mathbf{w}) - t_n\big)^2 + 2\sum_{i=1}^{D} w_i\mathbb{E}[\epsilon_i^{(n)}]\big(y(x_n, \mathbf{w}) - t_n\big) + \mathbb{E}[\big(\sum_{i=1}^{D} w_i\epsilon_i^{(n)}\big)^2]\big]$
$\because \mathbb{E}[\epsilon_i] = 0$
$=> \frac{1}{2}\sum_{n=1}^{N}\big[\big(y(x_n, \mathbf{w}) - t_n\big)^2 + \mathbb{E}[\big(\sum_{i=1}^{D} w_i\epsilon_i^{(n)}\big)^2]\big]$
$\because \mathbb{E}[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$
$=> \frac{1}{2}\sum_{n=1}^{N}\big[\big(y(x_n, \mathbf{w}) - t_n\big)^2 + \sum_{i=1}^{D}\big(\sigma w_i\big)^2\big]$
$= \frac{1}{2}\sum_{n=1}^{N}\big(y(x_n, \mathbf{w}) - t_n\big)^2 + \frac{N}{2}\sigma^2\sum_{i=1}^{D} w_i^2$
$= \frac{1}{2}\sum_{n=1}^{N}\big(y(x_n, \mathbf{w}) - t_n\big)^2 + \lambda\sum_{i=1}^{D} w_i^2$

is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight -decay regularization term.

6

Let $\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$ where $\mathbf{B}$ is a diagonalizable matrix. λ are $\mathbf{A}$'s eigenvalues

Left Hand Side：

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}|$$
$$= \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\lambda_1\lambda_2\ldots\lambda_n)$$
$$= \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\lambda_1) + \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\lambda_2)+\ldots+\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\lambda_n)$$
$$= \frac{1}{\lambda_1}\frac{\mathrm{d}\lambda_1}{\mathrm{d}\alpha} + \frac{1}{\lambda_2}\frac{\mathrm{d}\lambda_2}{\mathrm{d}\alpha}+\ldots+\frac{1}{\lambda_n}\frac{\mathrm{d}\lambda_n}{\mathrm{d}\alpha}$$
$$= \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}$$

Right Hand Side:

$$\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)$$
$$= \mathrm{Tr}\left((\mathbf{P}^{-1}\mathbf{B}\mathbf{P})^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{P}^{-1}\mathbf{B}\mathbf{P})\right)$$
$$= \mathrm{Tr}\left(\mathbf{P}^{-1}\mathbf{B}^{-1}\mathbf{P}\left[\frac{\mathrm{d}\mathbf{P}^{-1}}{\mathrm{d}\alpha}\mathbf{B}\mathbf{P} + \mathbf{P}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha}\mathbf{P} + \mathbf{P}^{-1}\mathbf{B}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\alpha}\right]\right)$$
$$= \mathrm{Tr}\left(\mathbf{P}^{-1}\mathbf{B}^{-1}\mathbf{P}\frac{\mathrm{d}\mathbf{P}^{-1}}{\mathrm{d}\alpha}\mathbf{B}\mathbf{P} + \mathbf{P}^{-1}\mathbf{B}^{-1}\mathbf{P}\mathbf{P}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha}\mathbf{P} + \mathbf{P}^{-1}\mathbf{B}^{-1}\mathbf{P}\mathbf{P}^{-1}\mathbf{B}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\alpha}\right)$$
$$= \mathrm{Tr}\left(\mathbf{P}\frac{\mathrm{d}\mathbf{P}^{-1}}{\mathrm{d}\alpha} + \mathbf{B}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha} + \mathbf{P}^{-1}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\alpha}\right)$$
$$= \mathrm{Tr}\left(\frac{\mathrm{d}\mathbf{P}\mathbf{P}^{-1}}{\mathrm{d}\alpha} + \mathbf{B}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha}\right)$$
$$= \mathrm{Tr}\left(\mathbf{B}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha}\right)$$

$$\because \mathrm{Tr}(\mathbf{B}) = \mathrm{Tr}(\mathbf{P}\mathbf{A}\mathbf{P}^{-1}) = \mathrm{Tr}(\mathbf{A}) \quad \text{and} \quad \mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$$

$$\therefore \mathbf{B} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix} \quad \text{and} \quad \mathbf{B}^{-1} = \begin{pmatrix} \lambda_1^{-1} & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^{-1} & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \lambda_n^{-1} \end{pmatrix}$$

$$=> \mathrm{Tr}\left(\mathbf{B}^{-1}\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}\alpha}\right)$$
$$= \sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\lambda_\mathrm{i}}\frac{\mathrm{d}\lambda_\mathrm{i}}{\mathrm{d}\alpha}$$

So left hand side and right hand side are the same.