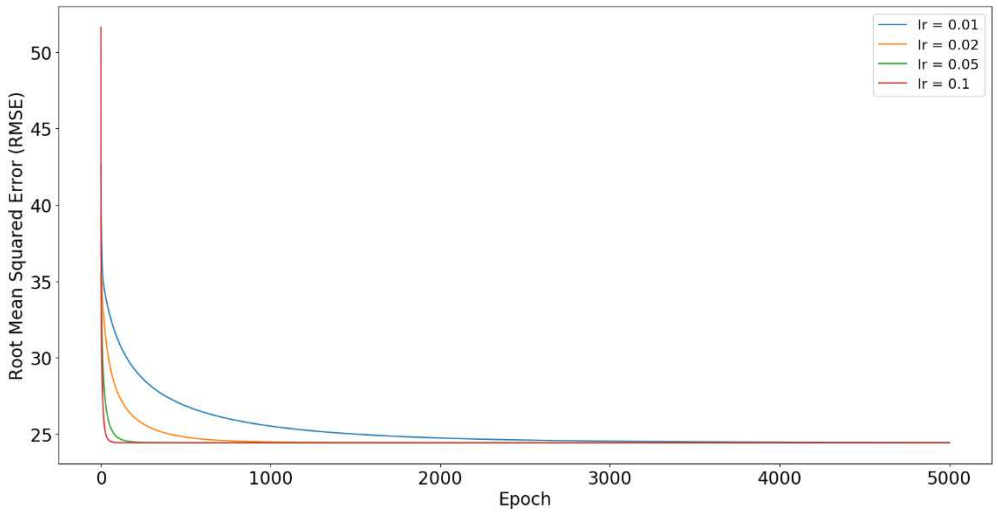
Homework 1 Report - PM2.5 Prediction

學號：D06922023　系級：資工所博士班二年級　姓名：顏志軒

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。



我分別使用 learning rate 0.01, 0.02, 0.05 以及 0.1 對於 PM2.5 的線性預測模型進行 5000 次梯度下降的迭代。由圖中可以看出，當 learning rate 越大時，RMSE 越快收斂到最後的值。而 learning rate 在這個例子中並不影響最後的結果。另外，一開始的 RMSE 在 learning rate = 0.1 時最大，可能是 learning rate 過大導致路徑偏離 loss 最小的點。

2. (1%) 請分別使用每筆 data 9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data 9 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）。

訓練模型	Training error	Testing error
Feature 包含全部參數	22.752870509876303	8.85828
Feature 僅有 PM2.5	23.3401409599984	9.62499

我分別使用過去 9 小時的所有參數(162 個參數)以及過去 9 小時的 PM2.5(9 個參數)進行訓練，所得到的 training error 與 testing error 如上表所示。由上表中可看出，包含全部參數的模型無論 training 或 testing error 都較低。這是因為包含的參數多，可以更容易貼近真正的模型，避免 under fitting 的情況。

另外，也可以用環境科學的 domain knowledge 解釋。風向、雨量都會影響一個地區 PM2.5 的濃度 [1]，在模型中排除這些因素，會影響到預測的準確度。

[1] 李亨山(2018)。4 工業測站 PM2.5 濃度增 環署：降雨減少。中央通訊社。網址：http://www.cna.com.tw/news/ahel/201808120118.aspx

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 parameter λ	Training error	Testing error
0	24.412036400415555	8.08303
0.01	24.415502114667227	8.0728
0.1	24.42228876758413	8.07502
10	24.5255159458738	8.203275
1000	29.694190264466968	10.051335

我在 b 與 W 初始值皆固定為 0 的情況下，以過去一小時的所有 18 個參數作為 feature，用五種不同的 regularization parameter 進行 training 以及 testing。由上表中可知，training error 隨著 regularization parameter 上升，而 testing error 則是微幅下降後也跟著上升。這個趨勢符合 regularization 可以讓 model 較為平滑，避免 overfitting training data 的結論。而 regularization 的效果並非特別顯著，可能是我原本就只有使用一小時的資料做 training，model 並不很複雜，因此 overfitting 的情況一開始就沒有很嚴重。

(數學題 4~6 請看下一頁)

4 Question 4

4.1 Question 4-a

Question Given tn is the data point of the data set $D = \{t_N, \ldots, t_N\}$. Each data point t_n is associated with a weighting factor $r_n > 0$. The sum-of-squares error function becomes:

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

Find the solution $\mathbf{w}^*$ that minimizes the error function.

Answer 假設 $\mathbf{w}$ 是一個 k 維的向量 $(w_1, w_2, \ldots, w_k)$。將 $E_D(\mathbf{w})$ 對 $\mathbf{w}$ 的第 i 個分量 w_i 偏微分得到:

$$\frac{\partial}{\partial w_i}E_D(\mathbf{w}) = \frac{1}{2}\cdot 2\sum_{n=1}^{N} r_n(t_n - \mathbf{w}^{\mathbf{T}}\mathbf{x}_n)(-x_{ni}) \tag{1}$$

方程式 1中，x_{ni} 表示向量 $\mathbf{x}_n$ 的第 i 個分量。
為求極值，令 $\frac{\partial}{\partial w_i}E_D(\mathbf{w})$ 在 w_i^* 的值為 0。因此:

$$0 = \sum_{n=1}^{N} r_n(t_n - \mathbf{w}^{*\mathbf{T}}\mathbf{x}_n)(-x_{ni}) \tag{2}$$

移項並取轉置得到:

$$\begin{aligned}\sum_{n=1}^{N} r_n t_n x_{ni} &= \sum_{n=1}^{N} r_n(\mathbf{w}^{*\mathbf{T}}\mathbf{x}_n)x_{ni} \\ &= \sum_{n=1}^{N} r_n x_{ni}(\mathbf{x}_n^{\mathbf{T}}\mathbf{w}^*)\end{aligned} \tag{3}$$

將所有分量合併回向量:

$$\sum_{n=1}^{N} r_n t_n \mathbf{x}_n = \sum_{n=1}^{N} r_n \mathbf{x}_n \mathbf{x}_n^{\mathbf{T}}\mathbf{w}^* \tag{4}$$

方程式 4中等號右邊的 $\mathbf{w}^*$ 與 n 無關，餘下的項為一個方陣。故:

$$\mathbf{w}^* = (\sum_{n=1}^{N} r_n \mathbf{x}_n \mathbf{x}_n^{\mathbf{T}})^{-1}\sum_{n=1}^{N} r_n t_n \mathbf{x}_n \tag{5}$$

$E_D(\mathbf{w})$ 對 w_i 二次偏微分的結果為:

$$\frac{\partial^2}{\partial w_i^2} E_D(\mathbf{w}) = \sum_{n=1}^{N} r_n x_{ni}^2 \tag{6}$$

此值為正，因此方程式 5中所列極值為最小值。

4.2 Question 4-b

Question Following the previous problem (4-a), if

$$\mathbf{t} = [t_1 t_2 t_3] = \begin{bmatrix} 0 & 10 & 5 \end{bmatrix}, \mathbf{X} = [\mathbf{x_1 x_2 x_3}] = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}$$

$$r_1 = 2, r_2 = 1, r_3 = 3$$

Answer 將數字代入

$$\begin{aligned} \sum_{n=1}^{N} r_n \mathbf{x}_n \mathbf{x}_n^{\mathbf{T}} &= 2 \begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix} + 1 \begin{bmatrix} 25 & 5 \\ 5 & 1 \end{bmatrix} + 3 \begin{bmatrix} 25 & 30 \\ 30 & 36 \end{bmatrix} = \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix} \\ \sum_{n=1}^{N} r_n t_n \mathbf{x}_n &= 0 \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 10 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + 15 \begin{bmatrix} 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 125 \\ 100 \end{bmatrix} \end{aligned} \tag{7}$$

故

$$\mathbf{w}^* = \frac{1}{2267} \begin{bmatrix} 127 & -107 \\ -107 & 108 \end{bmatrix} \begin{bmatrix} 125 \\ 100 \end{bmatrix} = \begin{bmatrix} \frac{5175}{2267} \\ -\frac{2575}{2267} \end{bmatrix} \tag{8}$$

5 Question 5

Question Given a linear model:

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

with a sum-of-squares error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \big(y(x_n, \mathbf{w}) - t_n)\big)^2$$

where t_n is the data point of the data set $\mathcal{D} = \{t_1, \ldots, t_N\}$

Suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ and $\mathbb{E}[\epsilon_i] = 0$ show that minimizing E averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Answer $E(\mathbf{w})$ 對於 ϵ_i 的平均為其期望值:

$$\begin{aligned}
\mathbb{E}[E(\mathbf{w})] =& \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}\left(w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_i) - t_n\right)^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}(w_0 - t_n)^2 + 2(w_0 - t_n)\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_i) + \left(\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_i)\right)^2\right] \\
=& \frac{1}{2}\left(\sum_{n=1}^{N}(w_0 - t_n)^2 + 2(w_0 - t_n)\sum_{i=1}^{D} w_i x_{ni} + \mathbb{E}\left[\left(\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_i)\right)^2\right]\right)
\end{aligned} \tag{9}$$

第三項進一步展開:

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_i)\right)^2\right] \\
=& \sum_{i=1}^{D}\left(w_i^2 x_{ni}^2 + w_i^2\mathbb{E}[\epsilon_i^2]\right) + 2\sum_{i=1}^{D}\sum_{\substack{j=1 \\ j\neq i}}^{D} w_i w_j \mathbb{E}[(x_{ni} + \epsilon_i)(x_{nj} + \epsilon_j)] \\
=& \sum_{i=1}^{D}\left(w_i^2 x_{ni}^2 + w_i^2\sigma^2\right) + 2\sum_{i=1}^{D}\sum_{\substack{j=1 \\ j\neq i}}^{D} w_i w_j x_{ni} x_{nj} \\
=& \left(\sum_{i=1}^{D} w_i x_{ni}\right)^2 + \sigma^2\sum_{i=1}^{D} w_i^2
\end{aligned} \tag{10}$$

合併回方程式 9，可以得到

$$
\begin{aligned}
&\mathbb{E}[E(\mathbf{w})] \\
=&\frac{1}{2}\left(\sum_{n=1}^{N}(w_0-t_n)^2+2(w_0-t_n)\sum_{i=1}^{D}w_i x_{ni}+\left(\sum_{i=1}^{D}w_i x_{ni}\right)^2+\sigma^2\sum_{i=1}^{D}w_i^2\right) \\
=&\frac{1}{2}\left(\sum_{n=1}^{N}\left(w_0-t_n+\sum_{i=1}^{D}w_i x_{ni}\right)^2+\sigma^2\sum_{i=1}^{D}w_i^2\right) \\
=&\frac{1}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w})-t_n)^2+\frac{1}{2}\sigma^2\sum_{i=1}^{D}w_i^2
\end{aligned}
\tag{11}
$$

方程式 11中，等號右邊的第一項為不包含高斯雜訊的 mean square error，因此在包含雜訊的 $\mathbb{E}[E(w)]$ 中對 $\mathbf{w}$ 最小化，相當於對於不包含雜訊的 $E(w)$ 做最小化，附帶一個 regularization parameter 為 $\frac{\sigma^2}{2}$ 的 regularization 項，且此項與 w_0 無關。

6 Question 6

Question $\mathbf{A}\in\mathbb{R}^{n\times n}, \alpha$, α is one of the elements of $\mathbf{A}$, prove that

$$
\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}|=\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)
$$

where the matrix $\mathbf{A}$ is a real, symmetric, non-singular matrix.

Answer

令 $\mathbf{T}$ 為 $n\times n$ 的實方陣，而 $\lambda_1,\lambda_2,\ldots,\lambda_n$ 為其 n 個特徵值。這 n 個數同時為以下兩個 n 次方程式的解：

$$
\begin{aligned}
\det(\mathbf{T}-\lambda\mathbf{I})&=0 \\
(\lambda_1-\lambda)(\lambda_2-\lambda)\ldots(\lambda_n-\lambda)&=0
\end{aligned}
\tag{12}
$$

兩式的 λ_n 項係數皆為 $(-1)^n$，故

$$
\det(\mathbf{T}-\lambda\mathbf{I})=(\lambda_1-\lambda)(\lambda_2-\lambda)\ldots(\lambda_{\mathrm{n}}-\lambda) \tag{13}
$$

將 $\lambda=-\frac{1}{\epsilon}$ 代入，得到：

$$\begin{aligned}\det(\mathbf{T}+\frac{1}{\epsilon}\mathbf{I}) &= (\lambda_1+\frac{1}{\epsilon})(\lambda_2+\frac{1}{\epsilon})\dots(\lambda_{\mathrm{n}}+\frac{1}{\epsilon})\\ \det(\mathbf{I}+\epsilon\mathbf{T}) &= (1+\epsilon\lambda_1)(1+\epsilon\lambda_2)\dots(1+\epsilon\lambda_{\mathrm{n}})\end{aligned} \tag{14}$$

ϵ 趨近於 0 時，等號右邊約為 $1+\epsilon\sum_{i=1}^{n}\lambda_i = \det(\mathbf{I})+\epsilon\mathrm{Tr}(\mathbf{T})$。取極限，得:

$$\lim_{\epsilon\to 0}\frac{\det(\mathbf{I}+\epsilon\mathbf{T})-\det(\mathbf{I})}{\epsilon} = \mathrm{Tr}(\mathbf{T}) \tag{15}$$

等號右邊為行列式的微分在 $\mathbf{I}$ 對 $\mathbf{T}$ 的方向導數。故

$$\det{}'(\mathbf{I})(\mathbf{T}) = \mathrm{Tr}(\mathbf{T}) \tag{16}$$

現在，考慮函數 $\det(\mathbf{X})$ 的微分。首先

$$\det(\mathbf{X}) = \det(\mathbf{A})\det(\mathbf{A}^{-1}\mathbf{X}) \tag{17}$$

根據連鎖律，其微分為:

$$\det{}'(\mathbf{X}) = \det(\mathbf{A})\det{}'(\mathbf{A}^{-1}\mathbf{X})\mathbf{A}^{-1} \tag{18}$$

將 $\mathbf{X}=\mathbf{A}$ 代入，得

$$\det{}'(\mathbf{A}) = \det(\mathbf{A})\det{}'(\mathbf{I})\mathbf{A}^{-1} \tag{19}$$

套用到 $\frac{d\mathbf{A}}{d\alpha}$ 上，得到

$$\begin{aligned}\det{}'(\mathbf{A})\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\alpha} &= \det(\mathbf{A})\det{}'(\mathbf{I})\mathbf{A}^{-1}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\alpha}\\ \frac{d}{d\alpha}\det(\mathbf{A}) &= \det(\mathbf{A})\mathrm{Tr}(\mathbf{A}^{-1}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\alpha})\\ Tr(\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha}) &= \frac{1}{\det(\mathbf{A})}\frac{d}{d\alpha}\det(\mathbf{A})\\ &= \frac{d}{d\alpha}\ln(\det(\mathbf{A}))\end{aligned} \tag{20}$$

即為所求。