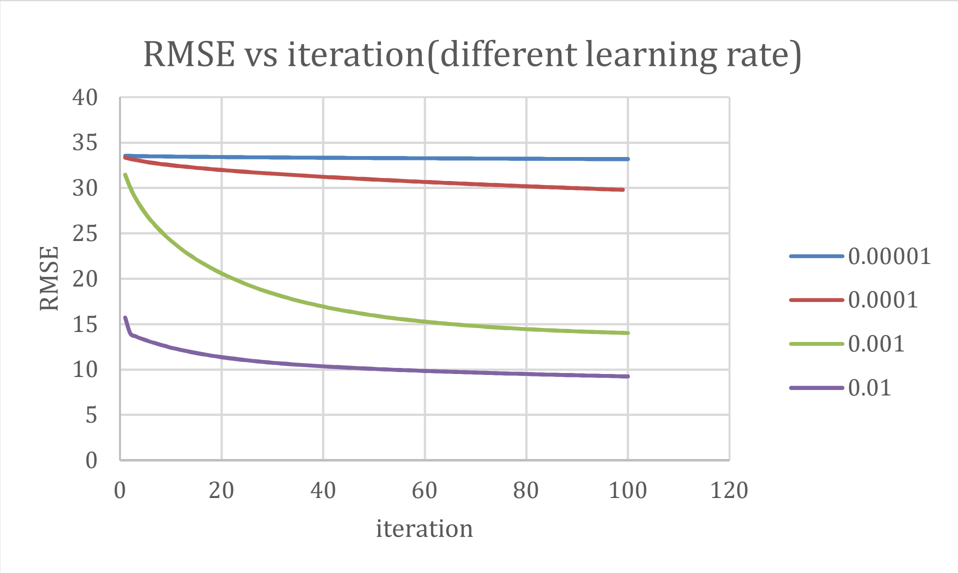
Homework 1 Report - PM2.5 Prediction

學號： R07522814 系級：機械所控制組碩一 姓名: 陳俊翰

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。



Learning rate 越大，收斂的速度越快且收斂值越小

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）。

	All feature	Pm2.5
Public score	8.60906	23.49060

只用 Pm2.5 會嚴重的 Underfit，可能原因是資料太少、曲線不夠平滑。

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。

λ	Training RMSE	Testing RMSE	Weigh 的 L2 norm
1	7.318612993441276	8.96643	0.47510398
10	7.320045123530217	8.96907	0.46528392
100	7.323611104232351	8.97510	0.44980106
1000	7.332340351665682	8.98838	0.41297003

λ 越大代表越 smooth，但是越大代表 loss function 考慮 training error 的比例降低，所以 training error 會增加，從上表可以看出，另外如果資料是越 smooth 的話 Testing error 應該會隨 λ 降低，但從我的程式中是增加的，最後 L2 norm 減小就和 Regurlization 的理論是一致的，Regurlization 就是為了讓 w 的總和越小。

4

4-a

Given t_n is the data point of the data set $\mathcal{D} = \{t_1, ..., t_N\}$. Each data point t_n is associated with a weighting factor $r_n > 0$.
The sum of squares error function becomes:

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

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

Ans:

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

$$\Rightarrow \sum_{n=1}^{N} r_n t_n x_n = \sum_{n=1}^{N} r_n(\mathbf{w}^T x_n)x(x_n) = \sum_{n=1}^{N} r_n(x_n^T w)x_n = \sum_{n=1}^{N} r_n x_n(x_n^T w)$$

$$\Rightarrow w^* = (\sum_{n=1}^{N} r_n x_n x_n^T)(\sum_{n=1}^{N} r_n t_n x_n)$$

#

4-b

Following the previous problem(2-a), if

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

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

Find the solution $\mathbf{w}^*$.

Ans:

$$\begin{aligned}\sum_{n=1}^{N} r_n x_n x_n^T &= 2\begin{bmatrix} 2 \\ 3 \end{bmatrix}[2 \quad 3] + 1\begin{bmatrix} 5 \\ 1 \end{bmatrix}[5 \quad 1] + 3\begin{bmatrix} 5 \\ 6 \end{bmatrix}[5 \quad 6] \\ &= \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}\end{aligned}$$

$$\Rightarrow (\sum_{n=1}^{N} r_n x_n x_n^T)^{-1} = \begin{bmatrix} 0.056 & -0.047 \\ -0.047 & 0.047 \end{bmatrix}$$

$$\sum_{n=1}^{N} r_n t_n x_n = 2 \times 0 \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 1 \times 10 \times \begin{bmatrix} 5 \\ 1 \end{bmatrix} + 3 \times 5 \times \begin{bmatrix} 5 \\ 6 \end{bmatrix}$$
$$= \begin{bmatrix} 125 \\ 100 \end{bmatrix}$$
$$\Rightarrow w^* = \begin{bmatrix} 0.0056 & -0.047 \\ -0.047 & 0.047 \end{bmatrix} \begin{bmatrix} 125 \\ 100 \end{bmatrix} = \begin{bmatrix} 2.28 \\ -1.13 \end{bmatrix}$$

#

5

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 $\mathcal{D} = \{t_1, ..., t_N\}$

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 $E[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ and $E[\epsilon_i] = 0$, 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.

Hint

$$\delta_{ij} = \begin{cases} 1(i = j), \\ 0(i \neq j). \end{cases} \tag{1}$$

Ans:

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

$$E^{'}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} [y^{'}(x, \mathbf{w}) - t_n]^2 = \frac{1}{2}(w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} w_i \epsilon_i - t_n)^2$$
$$= \frac{1}{2}\{[y(x, \mathbf{w}) - t_n)] + \sum_{i=1}^{D} w_i \epsilon_i\}^2$$

Thus, minimizing E averaged over the noise distribution is equivalent to minimizing the sum-of-squares error

6

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

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

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

Hint:The determinant and trace of $\mathbf{A}$ could be expressed in terms of its eigenvalues.

Ans:

Suppose $\mathbf{Au}$=$\lambda(\alpha)\mathbf{u}$, where $\lambda(\alpha)$ is eigenvalue and $\mathbf{u}$ is eigenvector

Because non-singular,

$$A^{-1}\mathbf{u} = \frac{1}{\lambda(\alpha)}\mathbf{u}$$

By Hint,

$$|A| = \prod \lambda_i \qquad Tr(\mathbf{A}) = \sum \lambda_i$$

Then,

$$\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha} = \mathbf{A}^{-1}\frac{d}{d\alpha}[\lambda(\alpha)\mathbf{u}] = \mathbf{A}^{-1}[\lambda^{'}(\alpha)\mathbf{u}] = \lambda^{'}(\alpha)\mathbf{A}^{-1}\mathbf{u} = \frac{\lambda^{'}(\alpha)}{\lambda(\alpha)}\mathbf{u}$$

$$\begin{aligned} Tr(\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha}) &= \sum \frac{\lambda_i^{'}(\alpha)}{\lambda_i(\alpha)} = \sum \frac{d}{d\alpha}[\ln \lambda_i(\alpha)] = \frac{d}{d\alpha}\ln(\prod \lambda_i(\alpha)) \\ &= \frac{d}{d\alpha}\ln|\mathbf{A}| \end{aligned}$$

QED