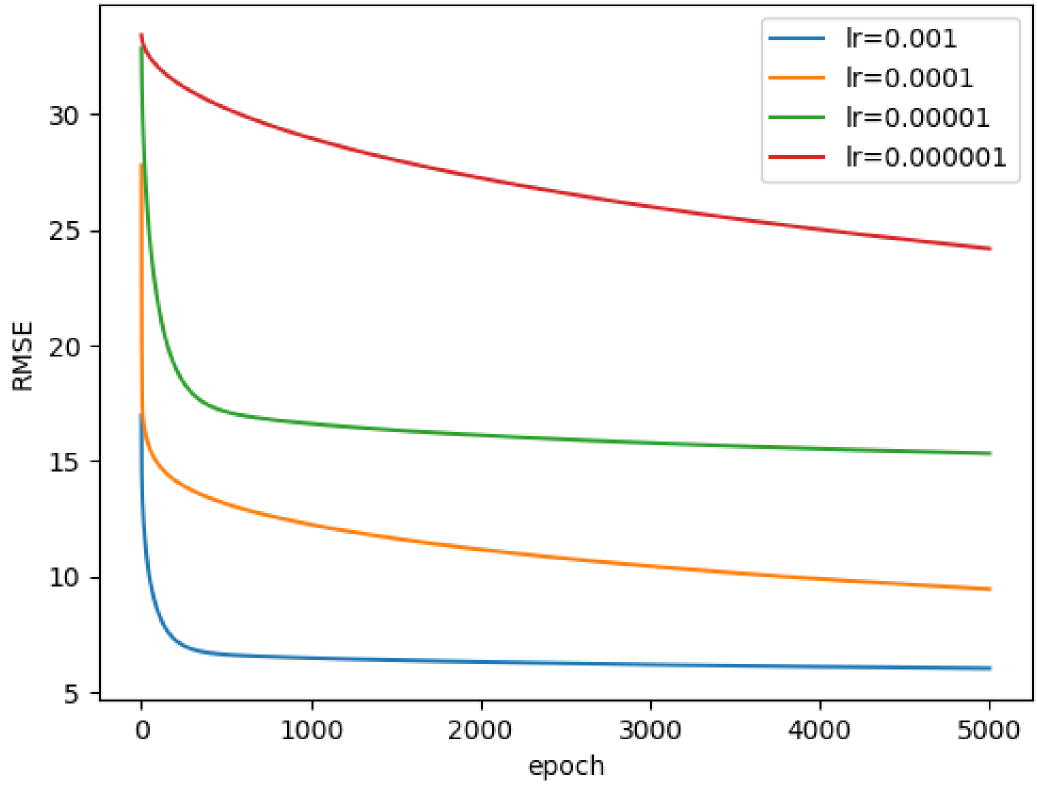
Homework 1 Report - PM2.5 Prediction

學號：R07943004　系級：電子所碩一　姓名: 莊育權

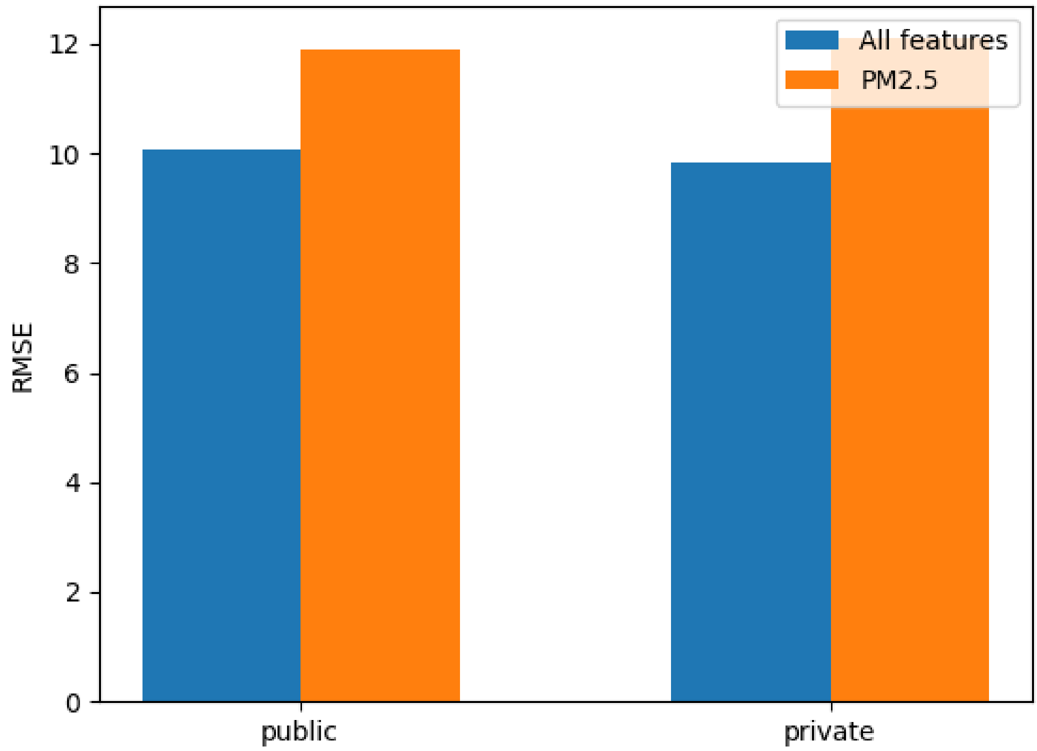
1. (1%) 請分別使用至少 4 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。



(batch size = 60, epoch = 5000, lamda = 1e-8, feature = PM2.5 和 PM2.5 平方, 有篩選資料)
由上圖可以看出，當 learning rate 越小，在初期的時候收斂速度會比較慢，而 learning rate 比較大的，在一開始就很快收斂，RMSE 的數值也很快就趨近於平緩。

No discussion with others

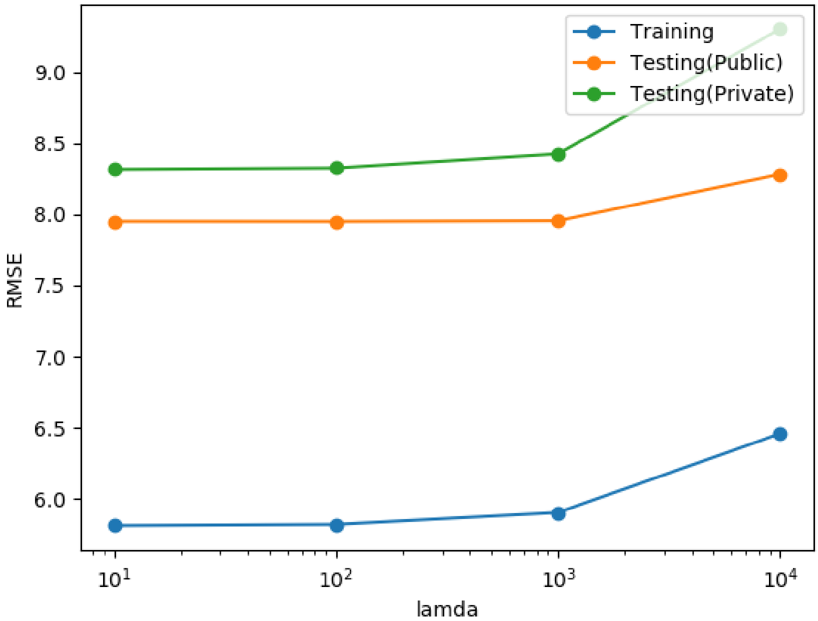
2. (1%) 請分別使用每筆 data 9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

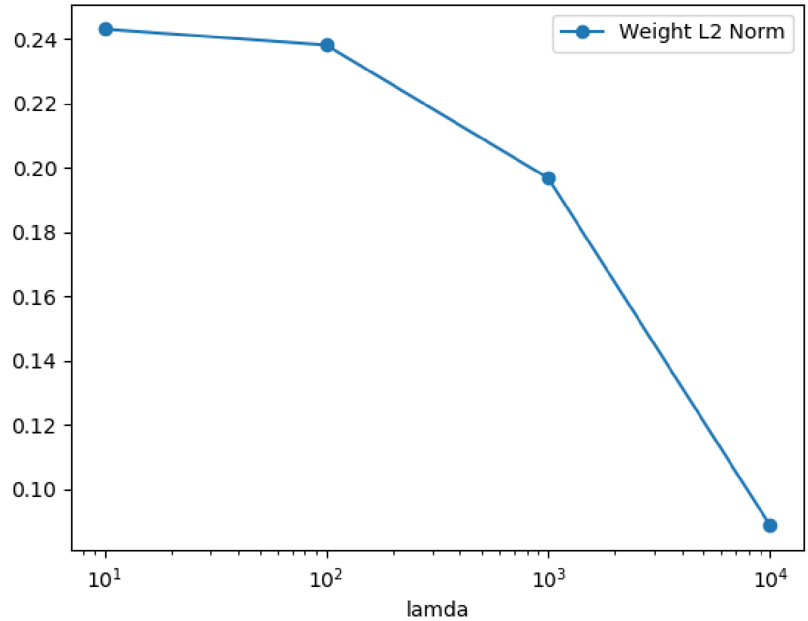


(batch size = 60, epoch = 10000, lr = 0.001, lamda = 1e-8, 沒有篩選資料)
由上圖可以觀察出，當全部的特徵都拿下去訓練，會比只拿 PM2.5 來的 RMSE 還要低。雖然可以知道並不是所有的特徵都對預測 PM2.5 有用，因此全部特徵拿下去訓練直覺上不一定會是好的，然而相對的，以此圖來說，可以知道，全部特徵裡面一定有幾項特徵也是跟 PM2.5 息息相關，因此 test 出來的 RMSE 才會比只拿 PM2.5 的還要低。

No discussion with others

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。





(batch size = 60, epoch = 10000, lr = 0.001, feature = PM2.5 和 PM2.5 平方, 有篩選資料)
在這次作業當中，由於特徵只取 PM2.5 和其平方，所以在項數並不多的情況下，小的正規化係數並不會造成太多的影響，如上圖，但當正規化係數過大的時候，會造成整體效能下降，可能是因為壓縮到權重自由的空間。

同時從下圖也可以看出，當正規化係數越大，的確權重會被限制。

No discussion with others

4. (1%)

(a)

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

Ans:

Let **R** is an NxN diagonal matrix, whose diagonal entries are $R_{i,i} = \frac{r_i}{2}$

$$E_D(\boldsymbol{w}) = (\boldsymbol{T} - \boldsymbol{X}\boldsymbol{w})^T \boldsymbol{R}(\boldsymbol{T} - \boldsymbol{X}\boldsymbol{w})$$
$$= (\boldsymbol{T}^T - \boldsymbol{w}^T\boldsymbol{X})\mathbf{R}(\mathbf{T} - \mathbf{X}\mathbf{w})$$
$$= \boldsymbol{T}^T\boldsymbol{R}\boldsymbol{T} - \boldsymbol{T}^T\boldsymbol{R}\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{T} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\boldsymbol{w}$$
$$\frac{\partial E_D(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0 - \boldsymbol{X}^T\boldsymbol{R}^T\boldsymbol{T} - \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{T} + 2\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\boldsymbol{w} = 0$$
$$\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{T}$$
$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{T}$$

No discussion with others

Ref:

https://onlinecourses.science.psu.edu/stat501/node/352/

https://math.stackexchange.com/questions/756679/least-squares-residual-sum-of-squares-in-closed-form

(b)

Following the previous problem(2-a), if

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

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

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

Ans:

$$\text{Let } \mathbf{X} = \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix}, \boldsymbol{R} = \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \boldsymbol{T} = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$
$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{T}$$

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

$$= \left(\frac{1}{2} \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix} \right)^{-1} \begin{bmatrix} 62.5 \\ 50 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{5175}{2267} \\ -\frac{2575}{2267} \end{bmatrix} \approx \begin{bmatrix} 2.283 \\ -1.136 \end{bmatrix}$$

No discussion with others

Ref:

https://matrixcalc.org/zh/#%7B%7B127/2267,-107/2267%7D,%7B-107/2267,108/2267%7D%7D%2A%7B%7B125%7D,%7B100%7D%7D

5. (1%)

Given a linear model:

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

with a sum-of-squares error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(\boldsymbol{x}_n, \mathbf{w}) - t_n\right)^2$$

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.

Hint

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

Ans:

$$\text{Let } y^{noisy} = \sum_i w_i x_i + \sum_i w_i \varepsilon_i \text{, } where\ \varepsilon_i\ is\ sampled\ from\ N(0, \sigma^2)$$

$$\mathrm{E}\left[\left(y^{noisy} - \mathrm{t}\right)^2\right] = \mathrm{E}\left[\left(y + \sum_i w_i \varepsilon_i - \mathrm{t}\right)^2\right] = \mathrm{E}\left[\left((y - \mathrm{t}) + \sum_i w_i \varepsilon_i\right)^2\right]$$

$$= (y - t)^2 + \mathrm{E}\left[2(y - \mathrm{t}) \sum_i w_i \varepsilon_i\right] + \mathrm{E}\left[\left(\sum_i w_i \varepsilon_i\right)^2\right]$$

$$Because\ \varepsilon_i\ is\ independent\ of\ \varepsilon_j\ and\ \varepsilon_i\ is\ independent\ of\ (y - t)$$

$$= (y - t)^2 + \mathrm{E}\left[\sum_i {w_i}^2 {\varepsilon_i}^2\right]$$

$$= (y - t)^2 + \sigma^2 \sum_i {w_i}^2$$

No discussion with others

Ref: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec9.pdf

6. (1%)

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

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

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

Hint:

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

Ans:

By using Chain Rule, $\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\det\mathbf{A}) = \frac{\mathrm{d}\ln(\det\mathbf{A})}{\mathrm{d}(\det\mathbf{A})}\frac{\mathrm{d}(\det\mathbf{A})}{\mathrm{d}\alpha}$

According to Jacobi's formula, $\frac{\mathrm{d}(\det\mathbf{A})}{\mathrm{d}\alpha} = \det(\mathbf{A})\,\mathrm{Tr}\left(\mathrm{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)$

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

No discussion with others

Ref: https://en.wikipedia.org/wiki/Jacobi%27s_formula