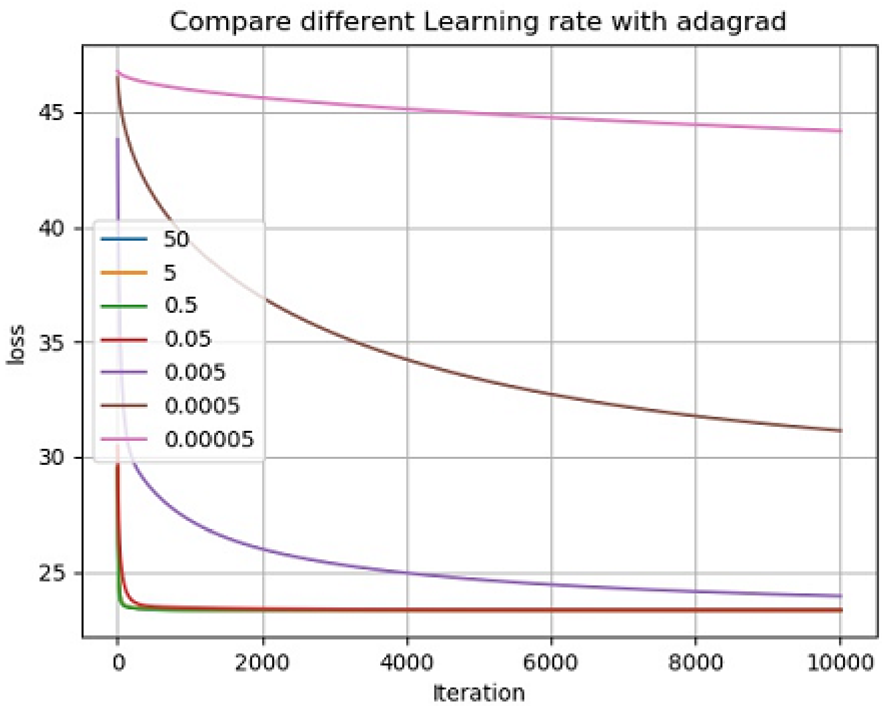
Homework 1 Report - PM2.5 Prediction

學號：B05505004 系級：工海三 姓名:朱信靂

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	50	5	0.5	0.05	0.005	0.0005
第0次iteration	18061	1744	148	35	45	46
第1次iteration	46	46	46	42	31	44
第2次iteration	29	29	29	29	30	43

這個實驗是比較只取9個小時PM2.5的model在不同的Learning rate得差別。

由於呈現方便，我把0~2個iteration的RMSE數據獨立出來改成用表格呈現，根據adagrad的公式，事實上w_0會等於$-(learning\ rate)$，也就是說Learning rate越大就會使得$|w_0|$越大，連帶使得第0次iteration RMSE值越大。但這邊就可以看到adagrad強大之處，他馬上把所有的值在第1次iteration就拉到差不多的數值。然而，可以看到Learning rate < 0.5的model原先在w_0數值就很小，可是卻因為Learning rate很小的關係，反而更新得很慢，導致iteration要設很大才會收斂。

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）。

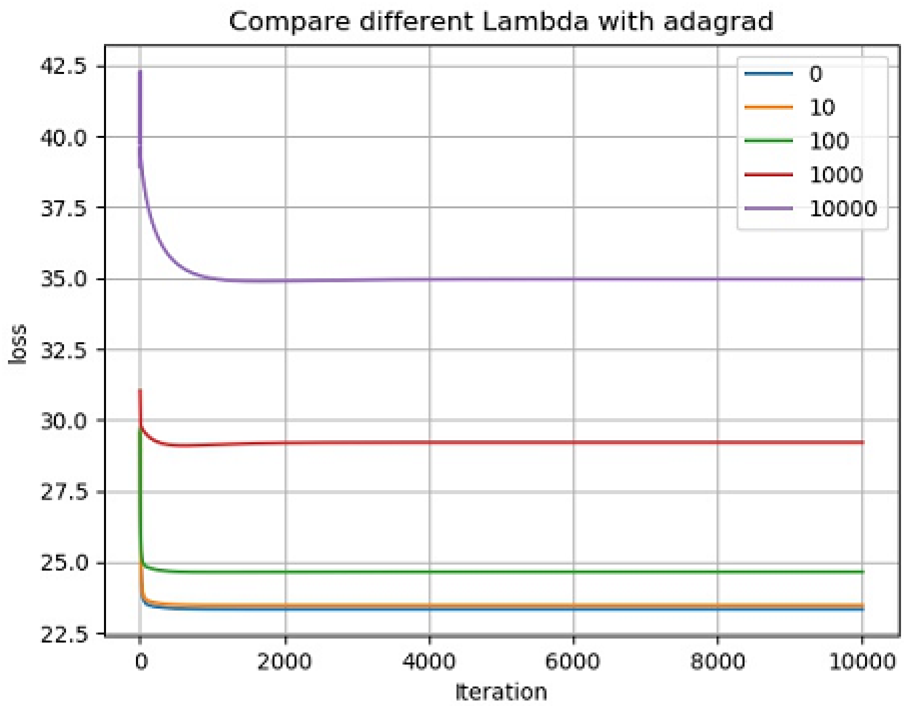
	Training Loss	Validation Loss	Public Score	Private Score
All feature	22.73722	14.98596	8.77603	8.35135
Only PM2.5	23.34014	9.04079	8.93422	9.06735

除了取的feature不一樣外，這個model的learning rate是0.9，總共是用50000 periods下去訓練，另外我是把0.1%的training data當作Validation Set。

1. 首先從上面的數據我們可以注意到Training Loss的值都遠大於其他三組的數據，不論feature取多還是少。換句話說我們可以知道training data上的數據一定存在著蠻大量的誤差，而這也可以從人工檢查上發現，照理來說資料除了每月的20號外的資料都應該要是連續的，可是卻很常會出現突然為0，甚至是負的數值的狀況。
2. 再來是feature取全部的model，雖然Validation的loss較高，可是performance在Public以及Private都比只取PM2.5的model好。也就是說有可能其他feature有PM2.5沒有的資訊，但這一點還必須再繼續實驗才可以驗證。
3. 另外還有一點比較特別的是只取PM2.5的model在Public取得較Private好的成績，而取全部feature的則是相反。這點也代表Public 和Private的data分布並非完全一樣，兩者還是有存在差異，而這個差異也導致我的model，有點過於fit在Public data上，而在Private的data表現很差。

3 (1%)

- 請分別使用至少四種不同數值的regularization parameter λ進行training（其他參數需一至），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。



Lambda	Training Loss	Validation Loss	Public Score	Private Score
0	23.34014	10.59402	9.55812	9.69186
10	23.47591	10.82571	9.63206	9.76802
100	24.65530	12.09714	10.32012	10.49241
1000	29.20624	15.18614	13.40702	13.68744
10000	34.96611	18.13653	18.80262	19.11563

在這個實驗中我是單純用9個小時的PM2.5當作feature來train model，由於這個model並沒有加入二次項，他只是單純的線性model，可以看到Loss隨著regularization parameter變大也跟著上升，並且在訓練的圖中更早就收斂，Loss無法繼續下降，產生嚴重的underfitting。

Lambda	Training Loss	Validation Loss	Public Score	Private Score
0	23.25445	12.44575	11.05005	11.52988
10	23.29365	12.51986	11.09599	11.56598

除此之外我還想比較若加入regularization在二次項的model是否會有比較好的performance，可以看到上面的數據雖然Training Loss都有下降，可是在Public及Private 都有蠻大的上升，明顯的overfit。可是可以看到在二次項若有加regularization，Public及Private的score上升幅度就沒有較一次的那麼快，也就是說regularization對於二次的model有較高的影響力，並且有解決到些許overfitting的問題。

4(1%)

(4-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:

  Set $\mathbf{w}, \mathbf{t}$ are column vector, and $\mathbf{R}$ is the diagonal matrix of weights $\mathbf{r}$

$$\mathbf{R} = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n \end{bmatrix}$$

Our goal is to find $\mathbf{w}^* = \arg\min E_D(\mathbf{w})$

$$
\begin{aligned}
E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N} r_n(t_n - \mathbf{w^T}\mathbf{x}_n)^2 \\
&= \frac{1}{2}(t - X^T w)^T R(t - X^T w) \\
&= \frac{1}{2}(t^T - w^T X)R(t - X^T w) \\
&= \frac{1}{2}(t^T Rt - t^T RX^T w - w^T XRt + w^T XRX^T w) && (1) \\
&\because E(w + \Delta w) - E(w) = \nabla_w E(W) \cdot \Delta w && (2)
\end{aligned}
$$

by (1)

$$
\begin{aligned}
E(w + \Delta w) - E(w) &= \frac{1}{2}(t^T Rt - t^T RX^T(w + \Delta w) - (w + \Delta w)^T XRt + (w + \Delta w)^T XRX^T(w + \Delta w)) \\
&- \frac{1}{2}(t^T Rt - t^T RX^T w - w^T XRt + w^T XRX^T w) \\
&= \frac{1}{2}(\Delta w^T XRX^T \Delta w - t^T RX^T \Delta w - \Delta w^T XRt + w^T XRX^T \Delta w + \Delta w^T XRX^T w)
\end{aligned}
$$

Since $scalar^T = scalar$

$$
\begin{gathered}
-t^T RX^T \Delta w = -\Delta w^T XRt \\
w^T XRX^T \Delta w = \Delta w^T XRX^T w
\end{gathered}
$$

$$
\begin{aligned}
E(w + \Delta w) - E(w) &= \frac{1}{2}(\Delta w^T XRX^T \Delta w - \Delta w^T XRt - \Delta w^T XRt + \Delta w^T XRX^T w + \Delta w^T XRX^T w) \\
&= \frac{1}{2}[\Delta w^T(2XRX^T w - 2XRt + XRX^T \Delta w)] \\
&= (XRX^T w - XRt + \frac{1}{2}XRX^T \Delta w)\Delta w^T
\end{aligned}
$$

since $\Delta w \to 0$ and from (2) we know that

$$\nabla_w E(W) = XRX^T w - XRt$$

To minimize $E_D(\mathbf{w})$, we set $\nabla_w E(W) = 0$, and then we get

$$\mathbf{w^*} = (XRX^T)^{-1}XRt \qquad (3)$$

(4-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:

$$\mathbf{R} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

from (3)

$$
\begin{aligned}
\mathbf{w}^* &= (\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix})^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} 5(1 \\
&= \begin{bmatrix} 0.056021 & -0.047199 \\ -0.047199 & 0.047640 \end{bmatrix} \begin{bmatrix} 125 \\ 100 \end{bmatrix} = \begin{bmatrix} 2.2828 \\ -1.1359 \end{bmatrix}
\end{aligned}
$$

5 (1%)

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

  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:

If we add the Gaussian noise to our model, the new model

$$
\begin{aligned}
y'(x_n, w) &= w_0 + \sum_{i=1}^{D} w_{ni}(x_{ni} + \epsilon_{ni}) \\
&= w_0 + \sum_{i=1}^{D} w_{ni} x_{ni} + \sum_{i=1}^{D} w_{ni} \epsilon_{ni} \\
&= y(x_n, w) + \sum_{i=1}^{D} w_{ni} \epsilon_{ni}
\end{aligned} \tag{4}
$$

According to (4) the new error function become

$$
\begin{aligned}
E'(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^{N} \left((y(x_n, w) - t_n) + \sum_{i=1}^{D} w_{ni} \epsilon_{ni})\right)^2 \\
&= \frac{1}{2} \sum_{n=1}^{N} \left((y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n) \sum_{i=1}^{D} w_{ni} \epsilon_{ni} + (\sum_{i=1}^{D} w_{ni} \epsilon_{ni})^2)\right)^2
\end{aligned} \tag{5}
$$

If we take expectation of equation (5)

$$
\mathbb{E}[E'(\mathbf{w})] = \frac{1}{2} \sum_{n=1}^{N} \left((y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n) \sum_{i=1}^{D} w_{ni} \mathbb{E}[\epsilon_{ni}] + \mathbb{E}[(\sum_{i=1}^{D} w_{ni} \epsilon_{ni})^2)\right)^2] \tag{6}
$$

since $\mathbb{E}[\epsilon_i] = 0$ and

$$
\begin{aligned}
\mathbb{E}[(\sum_{i=1}^{D} w_{ni}\epsilon_{ni})^2] = \mathbb{E}[\sum_{i=1}^{D}\sum_{i'=1}^{D} w_{ni}\epsilon_{ni}w_{ni'}\epsilon_{ni'}] = \sum_{i=1}^{D}\sum_{i'=1}^{D} w_{ni}w_{ni'}\mathbb{E}[\epsilon_{ni}\epsilon_{ni'}] \\
= \sum_{i=1}^{D}\sum_{i'=1}^{D} w_{ni}w_{ni'}\delta_{ii'}\sigma^2 = \sigma^2\sum_{i=1}^{D} w_i^2
\end{aligned}
$$

equation (6) become

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

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

6 (1%)

- $\mathbf{A} \in \mathbb{R}^{n\times n}, \alpha$ is one of the elements of $\mathbf{A}$, prove that $\frac{\mathrm{d}}{\mathrm{d}\alpha}ln|\mathbf{A}| = Tr\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)$

  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:

  since

$$
\det(I+\epsilon X) = 1 + Tr(X)\epsilon + O(\epsilon^2) \tag{7}
$$

By (4), therefore

$$
\begin{aligned}
\nabla_T|A| &= \lim_{\epsilon\to 0}\frac{\det(A+\epsilon T)-\det(A)}{\epsilon} = \lim_{\epsilon\to 0}\det(A)\frac{\det(I+\epsilon A^{-1}T)-\det(I)}{\epsilon} \\
&= \det(A)Tr(A^{-1}T)
\end{aligned}
$$

According to chain rule and the definition of directional derivative

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}ln|\mathbf{A}| &= \frac{\mathrm{d}ln|A|}{\mathrm{d}|A|}\cdot\frac{\mathrm{d}|A|}{\mathrm{d}A}\cdot\frac{\mathrm{d}A}{\mathrm{d}\alpha} = \frac{1}{|A|}\cdot\nabla_{\frac{\mathrm{d}A}{\mathrm{d}\alpha}}|A| \\
&= \frac{1}{|A|}\cdot\det(A)Tr(A^{-1}\frac{\mathrm{d}A}{\mathrm{d}\alpha}) = Tr(A^{-1}\frac{\mathrm{d}A}{\mathrm{d}\alpha})
\end{aligned}
$$

7(不算分，自行嘗試)

- 在第6中，若 $\mathbf{A}$不為symmetric，亦可推導出類似形式關係，可嘗試證明general case的推導，此部分不算分。