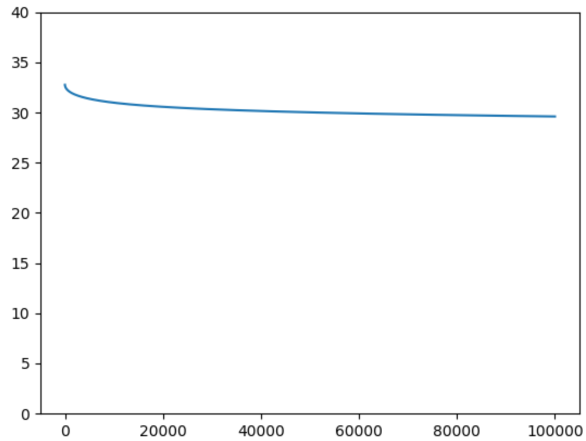
Homework 1 Report - PM2.5 Prediction

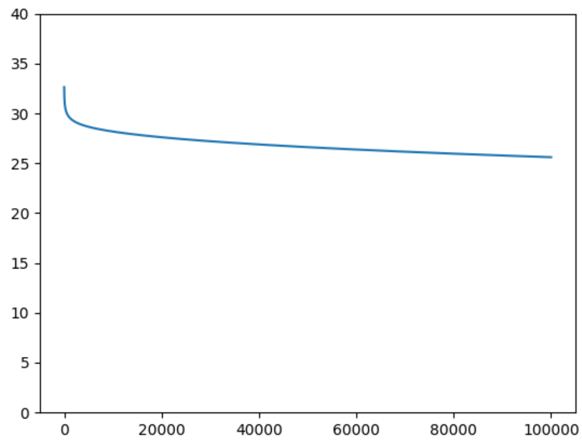
學號： B05902074 系級：資工三 姓名:魏佑珊

- Report.pdf 檔名錯誤（-1%）
- 學號系級姓名錯誤（-0.5%）

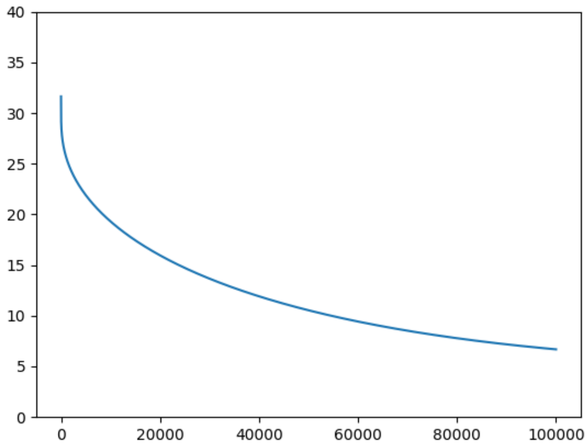
1. (1%) 請分別使用至少 **4** 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。
   (using ada grad, x: # of iteration, y: rmse)



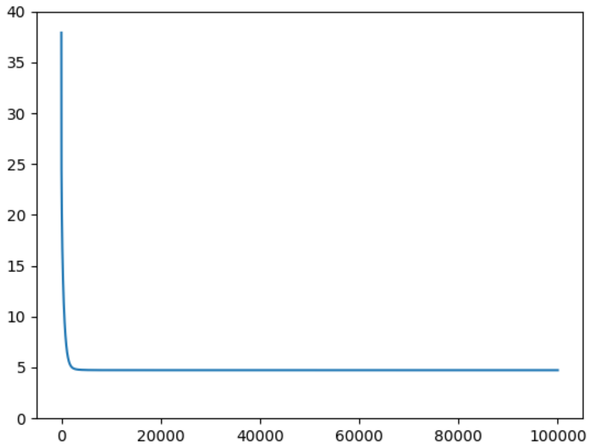
learning rate=0.0005



learning rate = 0.005



learning rate = 0.05



learning rate = 0.5

由圖可知，learning rate 越小，rmse 就越收斂得越慢。Learning rate = 0.5 時，在 iteration = 10000 之前就迅速收斂，之後 rmse 幾乎就沒甚麼改變了；但 learning rate = 0.0005 時，rmse 在 iteration = 20000~100000 仍持續下降，且速度緩慢。

會造成這個原因，是因為 learning rate 訂得較大，那麼每次 iteration 對 w 的修正也就比較大，使得 rmse 可以較快收斂；反之，若 learning rate 較小，收斂速度就會變慢。

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 RMSE	Public RMSE	Private RMSE
All feature	4.73515	6.28456	6.90698
Only PM2.5	5.35385	6.55759	7.02657

使用所有 feature 的 model，traning rmse 及 public leaderboard rmse 均小於只使用 PM2.5 train 出來的 model 對應的 rmse。由此推論，只使用 PM2.5 下去 train，feature 過少，會使得模型不能很好的去 fit data。因此，增加 feature 的數量才會使 rmse 有明顯的下降。

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。

	Training RMSE	Public RMSE	Private RMSE	Weight norm
Lambda = 5	4.73893	6.30551	6.88903	36.13085
Lambda = 10	4.74572	6.32357	6.86796	34.5558
Lambda = 20	4.76540	6.35968	6.83488	34.6980
Lambda = 30	4.78972	6.39625	6.81364	34.0519

以上分別使用 lambda = 5, 10, 15, 20 四種數值來 train model，原本我的預期是 training rmse 應該要隨著 lambda 上升而增加，反之 public 和 private score 應隨著 lambda 上升而減少，但不知為何 public 似乎也是增加的趨勢。不過以 training 和 private 的 rmse 評估的話，仍能觀察到 lambda 上升時，training error 也上升，但 testing error 卻下降。且 weight 的 norm 也呈現下降趨勢。這是因為 regularization term 的增加使得曲線更加平滑，weight 絕對值也較小，雖然不能很好的 fit training data，但曲線平滑度上升，所以有利於降低 testing data 的 error。

4.a

let X = $[\mathrm{x}_1 x_2 \ldots x_n]$

R =

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

Y = D^T

$\mathrm{E}_D(w)$

$= 1/2 * (\mathrm{R}*(\mathrm{X}^T w - Y)(X^T w - Y))$

$= 1/2 * (\mathrm{w}^T X - Y^T) R^T (X^T w - Y)$

$= 1/2 * (\mathrm{w}^T X R^T X^T w - w^T X R^T Y - Y^T R^T X^T w + Y^T R^T Y)$

$\nabla_w E_D(w) = XRX^T w - XRY$

w $=$ (XRX$^T)^{-1} XRY$

4.b

w $=$ (XRX$^T)^{-1} XRY =$

$$\left(\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} t1 \\ t2 \\ t3 \end{bmatrix}$$

$$= \begin{bmatrix} 2.28275254 \\ -1.13586237 \end{bmatrix}$$

5. Let

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} \tag{1}$$

and

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_D \end{bmatrix} \tag{2}$$

we have E(w) (with noise)

$= 1/2 \sum_{n=1}^{N} (w^T(x_n + \epsilon) - t_n)^2$

$= 1/2 \sum_{n=1}^{N} ((w^T x_n + w^T \epsilon) - t_n)((w^T x_n + w^T \epsilon) - t_n)$

$= 1/2 \sum_{n=1}^{N} ((w^T x_n - t_n)^2 + 2(w^T x_n - t_n) w^T \epsilon + w^T \epsilon w^T \epsilon)$

take the expected value of it:
$\mathbb{E}$(E(w))

$$=\mathbb{E}(1/2\sum_{n=1}^{N}(w^Tx_n-t_n)^2+1/2*2*\sum_{n=1}^{N}(w^Tx_n-t_n)w^T\epsilon+1/2\sum_{n=1}^{N}w^T\epsilon w^T\epsilon)$$

The second term should be zero, since after expansion, each term in it has a $\mathbb{E}(\epsilon_i)$, which is zero.
Now deal with the third term:

$$\mathbb{E}(1/2\sum_{n=1}^{N}w^T\epsilon w^T\epsilon)$$

$$=\mathbb{E}(1/2\sum_{n=1}^{N}(w_1\epsilon_1+w_2\epsilon_2+...+w_n\epsilon_n)(w_1\epsilon_1+w_2\epsilon_2+...+w_n\epsilon_n))$$

$$=\mathbb{E}(1/2\sum_{n=1}^{N}(\sum_{i=1}^{D}\sum_{j=1}^{D})w_iw_j\epsilon_i\epsilon_j)$$

Since $\mathbb{E}(\epsilon_i\epsilon_j)=\delta\sigma^2$, the above formula can be written as $\frac{N\sigma^2}{2}\sum_{i=1}^{D}w_i^2$

So, combine the first term and the third term, we get

$\mathbb{E}$(E(w))

$$=\mathbb{E}(1/2\sum_{n=1}^{N}(w^Tx_n-t_n)^2)+\frac{N\sigma^2}{2}\sum_{i=1}^{D}w_i^2$$

$$=(1/2\sum_{n=1}^{N}(w^Tx_n-t_n)^2)+\frac{N\sigma^2}{2}\sum_{i=1}^{D}w_i^2$$

which means the expected value of the E(w) with noise is same as the E(w) without noise add a weight-decay regularization term.
so minimizing the former is same as minimizing the latter.

**Discuss with B05902109 柯上優

6.
first prove det[exp(A)] = exp(Tr[A]):
A is a square matrix so we can have A = U $\cdot\Lambda \cdot U^{-1}$
where Λ is a diagonal matrix with the eigenvalues along its diagonal
and U are the corresponding eigenvectors.

we have f(A) = U $\cdot f(\Lambda) \cdot U^{-1}$

so det[f(A)] = det[u $\cdot f(\Lambda) \cdot U^{-1}$]

=det[U]det[f(Λ)]$\frac{1}{det[U]} = det[f(\Lambda)] = \prod_\alpha f(\Lambda_\alpha)$

for Trace:

Tr[f(A)] = Tr[U $\cdot f(\Lambda) \cdot U^{-1}$]

= Tr[U$^{-1} \cdot U \cdot f(\Lambda)$]
(Due to the cycle property of trace)

= Tr[f(Λ)] = $\sum_\alpha f(\Lambda_\alpha)$

so, det[exp(A)] = $\prod_\alpha exp(\Lambda_\alpha)$

= exp($\Lambda_1 + \Lambda_2 + ...$) $= exp(Tr[\Lambda]) = exp(Tr(A))$

using this property, we have

d/d$\alpha(ln(det[A])) = d/d\alpha(ln(det[exp(ln(A))]))$

= d/d$\alpha(ln(exp(Tr(ln(A))))) = d/d\alpha(Tr(ln(A)))$

=Tr(A$^{-1}\frac{d}{d\alpha}A$)

**Discuss with B05902083 余柏序