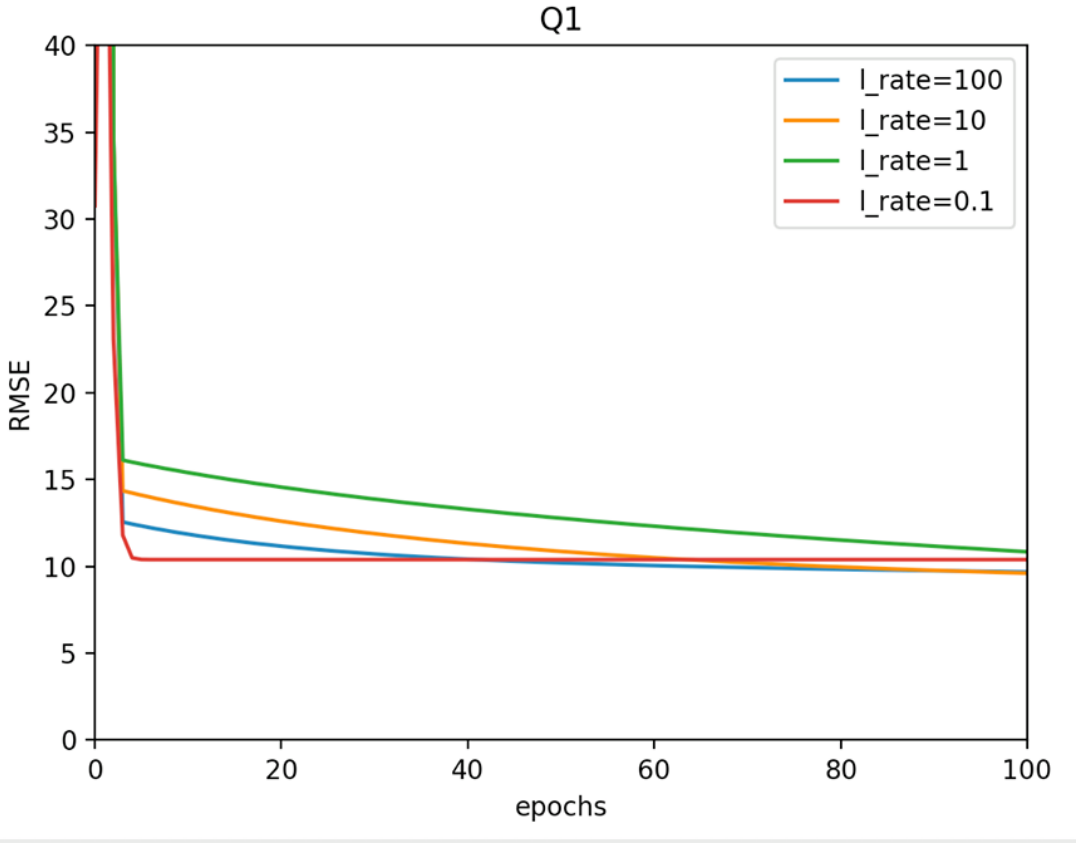
Homework 1 Report - PM2.5 Prediction

學號：R06942141　系級：電信所碩一　姓名: 詹鈞皓

- 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。



在 training 過程中，分別使用 100, 10, 1, 0.1 的 learning rate 來做訓練，根據上課所學，learning rate 就是控制要走的步伐大小，可以看到紅色線 learning rate 為 0.1，走的步伐最小，RMSE 轉折點位在 epochs 較高的地方，而其他的 learning rate 則在很早的 epoch 就轉折。而對於 RMSE 部分，較小的 learning rate 不一定有較好的表現，可能是較大的部分也可以走到谷底讓 RMSE 值最小。

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）。

	RMSE private	RMSE public
All feature	7.47687	7.00084
Only PM2.5	8.37752	7.12095

對於這兩個 models 在 public score 來看表現是差不多的，不過在 private score 來看，只使用 PM2.5 的 model 的表現是較差的，其原因可能是只使用 PM2.5 的特徵是不足以完整代表所有 data 的，在配對到未知測資表現就比較不好，而使用所有 feature 的 model 就可以比較全面地代表所有 data，但也有可能受到較不好的測資所影響結果。

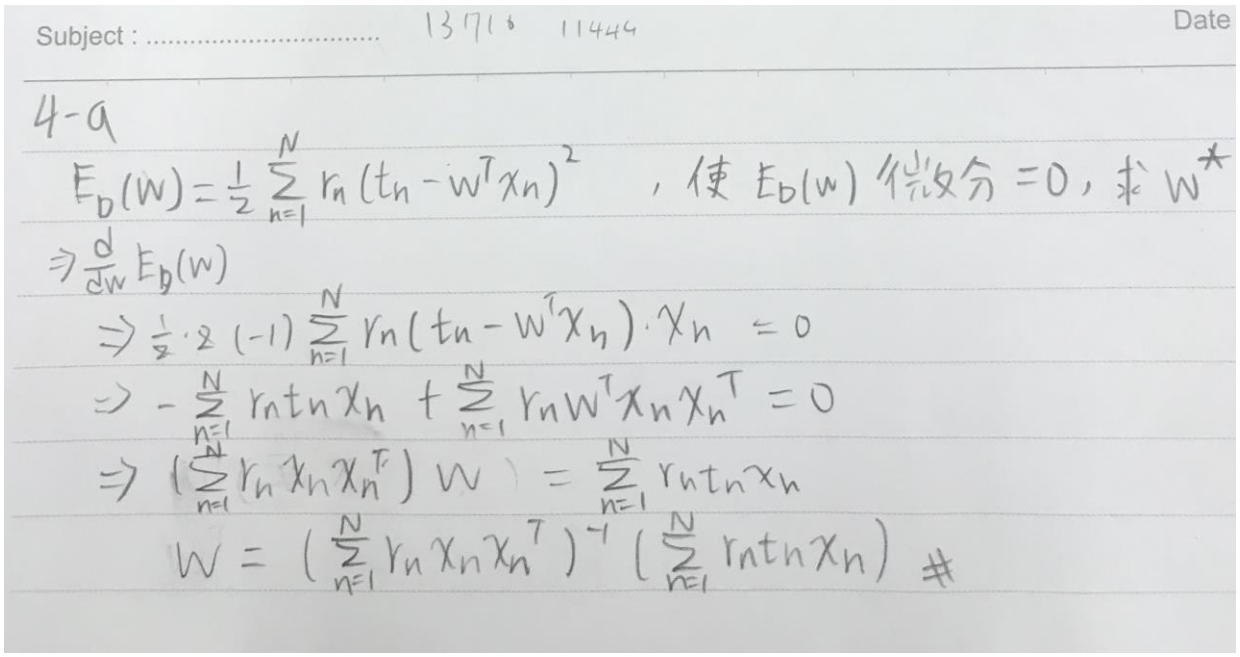
3. (1%)請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論及討論其 RMSE(traning, testing)（testing 根據 kaggle 上的 public/private score）以及參數 weight 的 L2 norm。

Regulization λ	RMSE private	RMSE public	L2 norm
1	8.17394	8.10156	0.28
10	10.14757	10.20079	0.089
100	12.55998	12.44503	0.04
1000	19.95427	19.98480	0.018

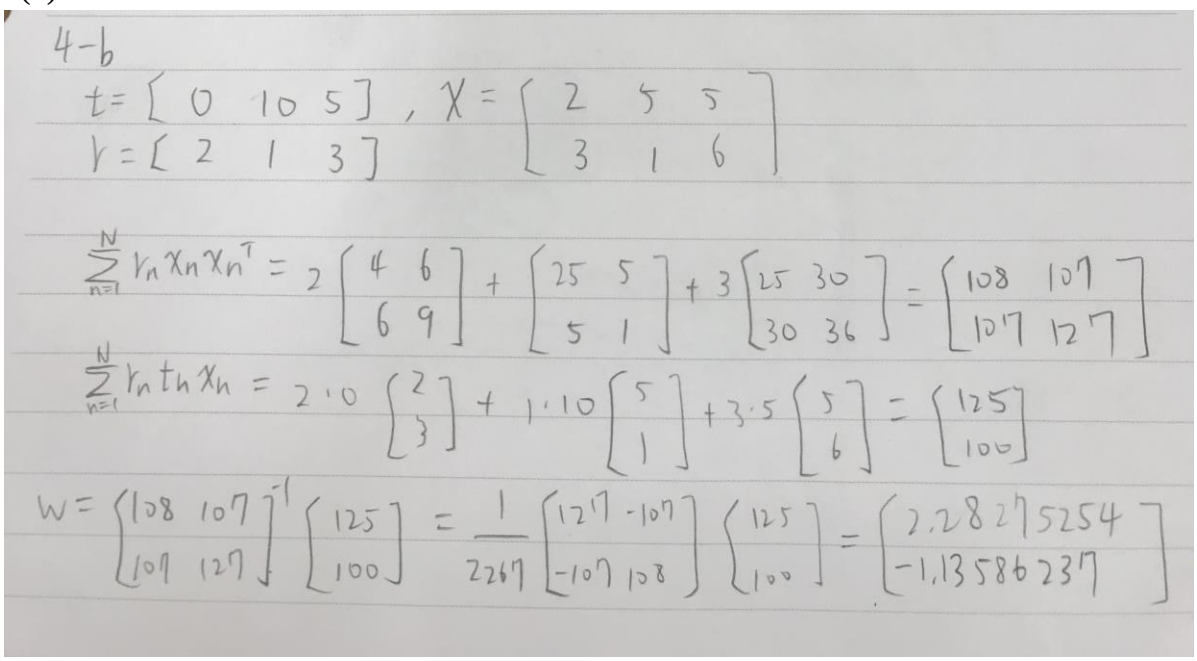
在 training 時，之所以要加入 regulization parameter 的原因是要避免選到太大的 W，因為選到太大的 W，對於誤差的影響也越大，因此需要這項 regulization parameter 來平衡，避免選到太大的 W。而在這個實驗中，可以發現如果用了太大的 λ 來做平衡反而會造成 RMSE 劇烈的攀升，所以在使用上也是要經過測試選取表現最好的參數作使用。

4~6 (3%) 請參考數學題目（連結：），將作答過程以各種形式（latex 尤佳）清楚地呈現在 pdf 檔中（手寫再拍照也可以，但請注意解析度）。

4(a).


Subject :
Date
4-a
$E_D(W)=\frac{1}{2}\sum_{n=1}^{N} r_n(t_n-W^T x_n)^2$ ，使 $E_D(W)$ 微分 $=0$，求 W^*
$\Rightarrow \frac{d}{dW}E_D(W)$
$\Rightarrow \frac{1}{2}\cdot 2(-1)\sum_{n=1}^{N} r_n(t_n-W^T x_n)\cdot x_n = 0$
$\Rightarrow -\sum_{n=1}^{N} r_n t_n x_n + \sum_{n=1}^{N} r_n W^T x_n x_n^T = 0$
$\Rightarrow (\sum_{n=1}^{N} r_n x_n x_n^T) W = \sum_{n=1}^{N} r_n t_n x_n$
$W=(\sum_{n=1}^{N} r_n x_n x_n^T)^{-1}(\sum_{n=1}^{N} r_n t_n x_n)$ #

4(b).


4-b
$t=[0\ \ 10\ \ 5]$, $X=\begin{bmatrix}2&5&5\\3&1&6\end{bmatrix}$
$r=[2\ \ 1\ \ 3]$
$\sum_{n=1}^{N} r_n x_n x_n^T = 2\begin{bmatrix}4&6\\6&9\end{bmatrix}+\begin{bmatrix}25&5\\5&1\end{bmatrix}+3\begin{bmatrix}25&30\\30&36\end{bmatrix}=\begin{bmatrix}108&107\\107&127\end{bmatrix}$
$\sum_{n=1}^{N} r_n t_n x_n = 2\cdot 0\begin{bmatrix}2\\3\end{bmatrix}+1\cdot 10\begin{bmatrix}5\\1\end{bmatrix}+3\cdot 5\begin{bmatrix}5\\6\end{bmatrix}=\begin{bmatrix}125\\100\end{bmatrix}$
$W=\begin{bmatrix}108&107\\107&127\end{bmatrix}^{-1}\begin{bmatrix}125\\100\end{bmatrix}=\frac{1}{2267}\begin{bmatrix}127&-107\\-107&108\end{bmatrix}\begin{bmatrix}125\\100\end{bmatrix}=\begin{bmatrix}2.282\,75254\\-1.135\,86237\end{bmatrix}$

5.

5.

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

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

加 λ noise

Let $\hat{y}_n = w_0 + \sum_{i=1}^{D} w_i (x_i^{(n)} + \varepsilon_i)$

$= w_0 + \sum_{i=1}^{D} (w_i x_i^{(n)} + w_i \varepsilon_i)$

令 $y'_n = w_0 + \sum_{i=1}^{D} w_i^{(n)} x_i$, y'_n is the Prediction from the current model for the n^{th} data point

$\therefore \hat{y}_n = y'_n + \sum_{i=1}^{D} w_i \varepsilon_i$

將 $\hat{y}_n$ 代入②式

$\Rightarrow E(w) = \frac{1}{2} \sum_{n=1}^{N} (\hat{y}_n - t_n)^2 = (\hat{y}_n^2 - 2\hat{y}_n t_n + t_n^2)$

$= \frac{1}{2} \sum_{n=1}^{N} (y'^2_n + 2y'_n \sum_{i=1}^{D} w_i \varepsilon_i + (\sum_{i=1}^{D} w_i \varepsilon_i)^2 - 2(y'_n + \sum_{i=1}^{D} w_i \varepsilon_i) t_n + t_n^2)$

$= \frac{1}{2} \sum_{n=1}^{N} (y'^2_n - 2y'_n t_n + t_n^2 + (\sum_{i=1}^{D} w_i \varepsilon_i)^2 + 2y'_n \sum_{i=1}^{D} w_i \varepsilon_i - 2 \sum_{i=1}^{D} w_i \varepsilon_i t_n)$

又 $\because E[\varepsilon_i] = 0$

$\therefore \sum_{i=1}^{D} w_i \varepsilon_i = 0$

$\therefore E(w) = \frac{1}{2} \sum_{n=1}^{N} (y'_n - t_n)^2 + (\sum_{i=1}^{D} w_i \varepsilon_i)^2$

又 $\because E[\varepsilon_i \varepsilon_j] = \delta_{ij} \sigma^2$, 當 $j = i$, $E[\varepsilon_i \varepsilon_j] = \sigma^2$

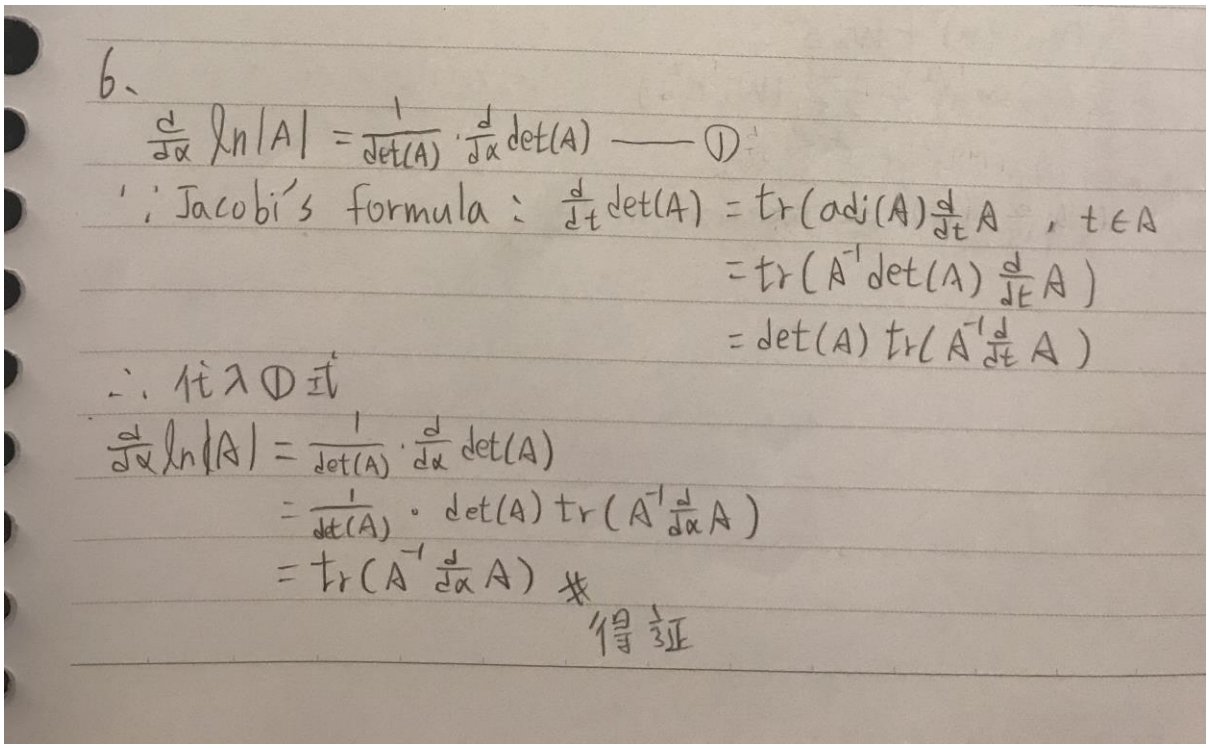
$\therefore E(w) = \frac{1}{2} \sum_{n=1}^{N} (y'_n - t_n)^2 + w_i \sigma^2$

$= \frac{1}{2} \sum_{n=1}^{N} (y'_n - t_n)^2 + \frac{1}{2} \sum_{n=1}^{N} (w_i \sigma^2)$

$= \frac{1}{2} \sum_{n=1}^{N} (y'_n - t_n)^2 + \frac{\sigma^2}{2} \sum_{i=1}^{N} w_i$

令 $\frac{\sigma^2}{2} = \lambda$, 則 $\lambda \sum_{i=1}^{N} w_i$ 為 regulization 項 ※

6.


6.
$\frac{d}{d\alpha}\ln|A| = \frac{1}{\det(A)}\cdot\frac{d}{d\alpha}\det(A)$ —— ①
∵ Jacobi's formula : $\frac{d}{dt}\det(A) = tr(adj(A)\frac{d}{dt}A$, $t\in A$
$= tr(A^{-1}\det(A)\frac{d}{dt}A)$
$= \det(A)\,tr(A^{-1}\frac{d}{dt}A)$
∴ 代入①式
$\frac{d}{d\alpha}\ln|A| = \frac{1}{\det(A)}\cdot\frac{d}{d\alpha}\det(A)$
$= \frac{1}{\det(A)}\cdot\det(A)\,tr(A^{-1}\frac{d}{d\alpha}A)$
$= tr(A^{-1}\frac{d}{d\alpha}A)$ ※
得証