Homework 1 Report - PM2.5 Prediction

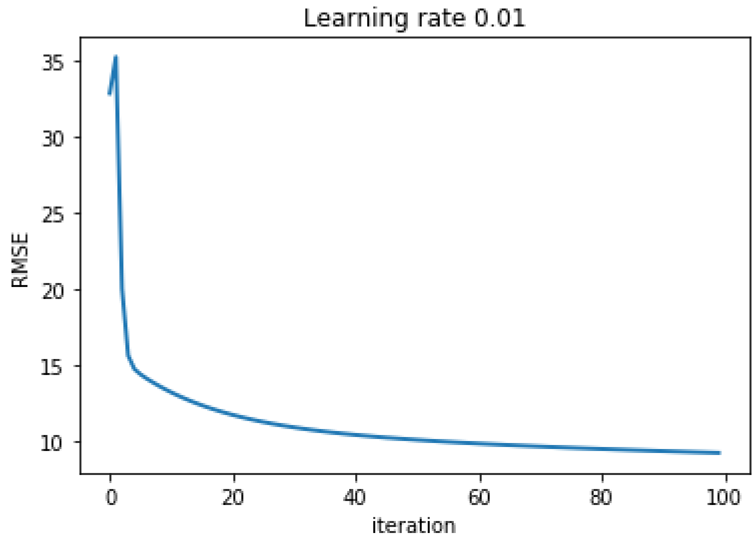
學號：R06922134　系級： 資工碩二 姓名: 葉沛陽

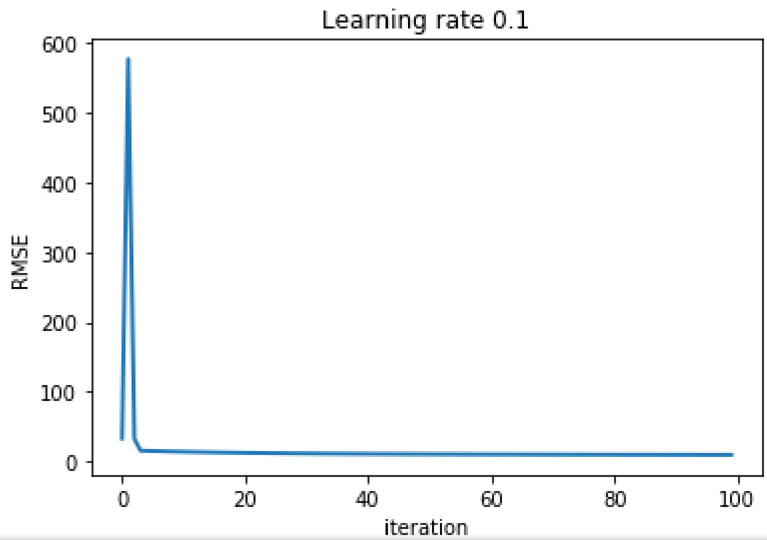
- Report.pdf 檔名錯誤（-1%）
- 學號系級姓名錯誤（-0.5%）

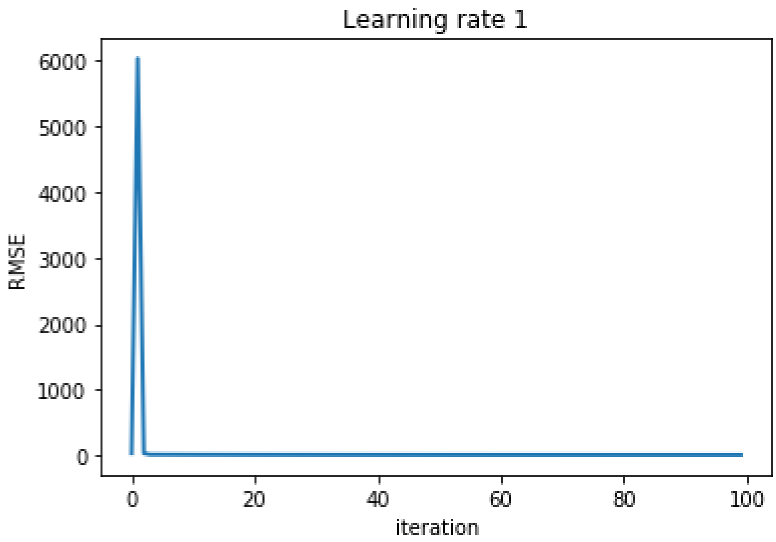
1. (1%) 請分別使用至少 **4** 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。

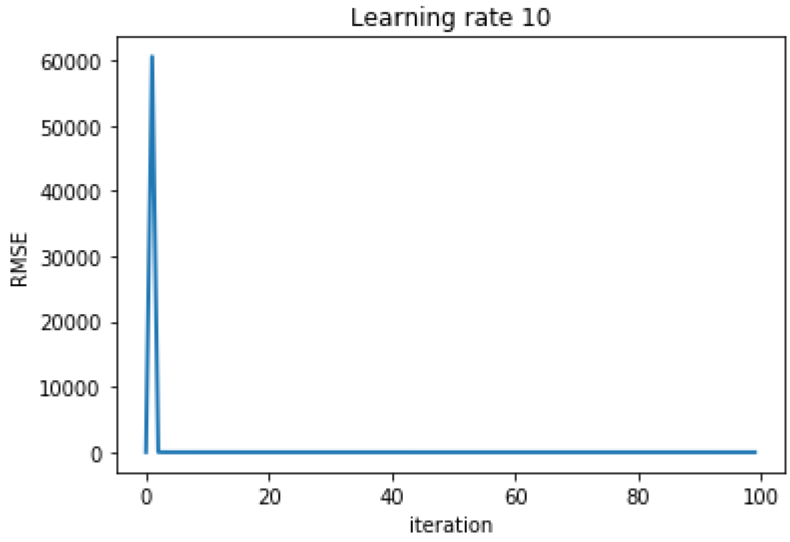
以下為示意圖：

Ans:









Learning rate 越大，在第 2 個 iteration 的 RMSE 會飆越高，不過 4 種 learning rate 在第 100 個 iteration 的時候，分別收斂到{learning rate 10 : RMSE:9.398166};{learning rate 1 : RMSE:9.391467};{learning rate 0.1 : RMSE:9.328023};{learning rate 0.01 : RMSE:9.215379};RMSE 都在 9 左右，只有些微差距，learning rate 越小 在第 100iteration 時 RMSE 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）。

	public	private
所有 feature	5.63946	6.81698
PM2.5	6.35282	7.40227

只用 PM2.5 雖然直觀上，PM2.5 是跟答案最相關的 feature，不過少了其他那麼多 feature 還是大大的影響了準確率，證明了其他 feature 也是跟 PM2.5 的變化有相關性的。

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。
Ans:

其他參數
Iteration = 10000 ;   Learning rate = 10 ;

λ	RMSE(training)	RMSE(public)	RMSE(private)	Weight L2 norm
0.01	22.603668	9.00778	8.73400	1.87981517
0.1	22.603668	9.00778	8.73400	1.87971463
1	22.603671	9.00776	8.73401	1.87870999
10	22.603703	9.00754	8.73402	1.86873505

regulization parameter λ 越大 RMSE(training)有稍微變大 不過 Weight L2 norm 會變小來調控，因此改變不大。
regulization parameter λ 越大 RMSE(public)有稍微變小 不過 Weight L2 norm 會變小來調控，因此改變不大。
regulization parameter λ 越大 RMSE(private)有稍微變大 不過 Weight L2 norm 會變小來調控，因此改變不大。
regulization parameter λ 越大 會讓 Weight 的改變對 Loss 的影響放大 所以 Weight L2 norm 會隨著 λ 變大而變小
然後可以發現 RMSE(training)遠比 RMSE(public)跟 RMSE(private)來得大，因為加入了 regulization 會讓 RMSE(training)變很大，所以這是當然的
至於一般來說加入 regulization 會比較不會 overfitting，不過這裡看不出來。

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.
(4-a)

Ans:

設$x_i,...,x_N$是d維。

$$E_D(\mathbf{w})=\frac{1}{2}(r_1(t_1-(w_1x_{1,1}+...+w_dx_{1,d}))^2+...+r_N(t_N-(w_1x_{N,1}+...+w_dx_{N,d}))^2)$$

$$\frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}_i}=(r_1(w_1x_{1,1}+...+w_dx_{1,d}-t_1)x_{1,i}+...+(r_N(w_1x_{N,1}+...+w_dx_{N,d}-t_N)x_{N,i})$$

$$\triangledown E_D(w)=\begin{bmatrix}\frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}_1}\\ \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}_2}\\ .\\ .\\ .\\ \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}_d}\end{bmatrix}=\begin{bmatrix}r_1x_1 & r_2x_2 & \dots & r_Nx_N\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ .\\ .\\ .\\ x_N\end{bmatrix}\begin{bmatrix}w_1\\ w_2\\ .\\ .\\ .\\ w_d\end{bmatrix}-\begin{bmatrix}r_1x_1 & r_2x_2 & \dots & r_Nx_N\end{bmatrix}\begin{bmatrix}t_1\\ t_2\\ .\\ .\\ .\\ t_N\end{bmatrix}$$

令

$$X'=\begin{bmatrix}r_1x_1\\ r_2x_2\\ .\\ .\\ .\\ r_Nx_N\end{bmatrix},X=\begin{bmatrix}x_1\\ x_2\\ .\\ .\\ .\\ x_N\end{bmatrix},W=\begin{bmatrix}w_1\\ w_2\\ .\\ .\\ .\\ w_d\end{bmatrix},T=\begin{bmatrix}t_1\\ t_2\\ .\\ .\\ .\\ t_N\end{bmatrix}$$

For

$$\triangledown E_D(W^*)=0$$

$$(X')^TXW^*=(X')^TT$$

$$W^*=((X')^TX)^{-1}(X')^TT$$

(4-b)

ANS:

$$W^*=((X')^TX)^{-1}(X')^T=(\begin{bmatrix}4 & 5 & 15\\ 6 & 1 & 18\end{bmatrix}\begin{bmatrix}2 & 3\\ 5 & 1\\ 5 & 6\end{bmatrix})^{-1}\begin{bmatrix}4 & 5 & 15\\ 6 & 1 & 18\end{bmatrix}\begin{bmatrix}0\\ 10\\ 5\end{bmatrix}$$

$$=\begin{bmatrix}108 & 107\\ 107 & 127\end{bmatrix}^{-1}\begin{bmatrix}125\\ 100\end{bmatrix}=\begin{bmatrix}\frac{127}{2267} & \frac{-107}{2267}\\ \frac{-107}{2267} & \frac{108}{2267}\end{bmatrix}\begin{bmatrix}125\\ 100\end{bmatrix}=\begin{bmatrix}\frac{5175}{2267}\\ \frac{-2575}{2267}\end{bmatrix}$$

5.

ANS:

$$E_{noise}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\big(y(x_n+\epsilon_i,\mathbf{w})-t_n\big)^2 = \frac{1}{2}\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n+\sum_{i=1}^{D}w_i\epsilon_i)^2$$

$$=\frac{1}{2}\sum_{n=1}^{N}((\sum_{i=1}^{D}w_ix_i+w_0-t_n)^2+\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n)(\sum_{i=1}^{D}w_i\epsilon_i)+\frac{1}{2}N(\sum_{i=1}^{D}w_i\epsilon_i)^2)$$

$$=\frac{1}{2}\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n)^2+\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n)(\sum_{i=1}^{D}w_i\epsilon_i)+\frac{1}{2}N(\sum_{i=1}^{D}w_i\epsilon_i)^2$$

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

$$=\frac{1}{2}\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n)^2+\frac{1}{2}N\sigma^2\sum_{i=1}^{D}w_i^2$$

let λ = $N\sigma^2$

$$=\frac{1}{2}\sum_{n=1}^{N}(\sum_{i=1}^{D}w_ix_i+w_0-t_n)^2+\frac{1}{2}\lambda\sum_{i=1}^{D}w_i^2$$

(the addition of a weight-decay regularization term)

6.

ANS:

$det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$
$trace(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$
$\mathbf{A}$ is a real, symmetric, non-sigular matrix. So, $\mathbf{A}$ is diagonalizable matrix.
$\mathbf{A} = P^{-1}DP$

if α is eigenvalue of $\mathbf{A}$, and there is h eigenvalues $\lambda_k,...,\lambda_{k+h}$ is equal to α ,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} ln(\prod_{i=1}^{n} \lambda_i) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{i=1}^{n} (ln\lambda_i) = \frac{1}{\lambda_k} + \frac{1}{\lambda_{k+1}} + \ldots + \frac{1}{\lambda_{k+h}}$$

$$\mathbf{A}^{-1} = P^{-1}D^{-1}P$$

let

$$D' = \begin{bmatrix} \frac{\partial\lambda_1}{\partial\alpha} & 0 & \ldots & & 0 \\ . & \frac{\partial\lambda_2}{\partial\alpha} & 0 & \ldots & 0 \\ . & & & & \\ . & & & & \\ 0 & 0 & \ldots & 0 & \frac{\partial\lambda_n}{\partial\alpha} \end{bmatrix}$$

$$Tr(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}) = P^{-1}D^{-1}P = Tr(P^{-1}D^{-1}PP^{-1}D'P) = Tr(D^{-1}D') = \frac{1}{\lambda_k} + \frac{1}{\lambda_{k+1}} + \ldots + \frac{1}{\lambda_{k+h}}$$

So,

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