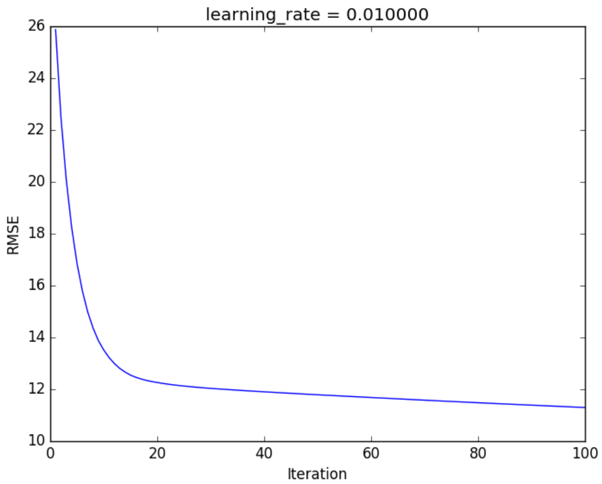
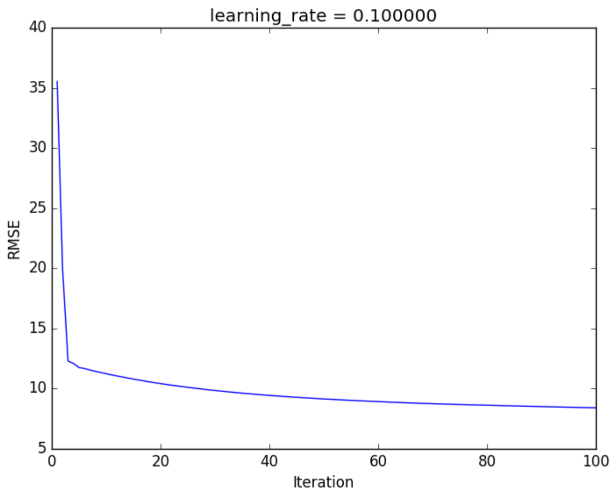
Homework 1 Report - PM2.5 Prediction

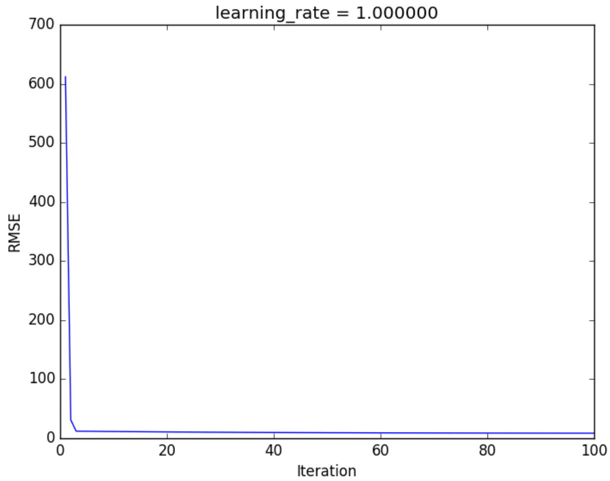
學號： B05902001 系級： 資工三 姓名:廖彥綸

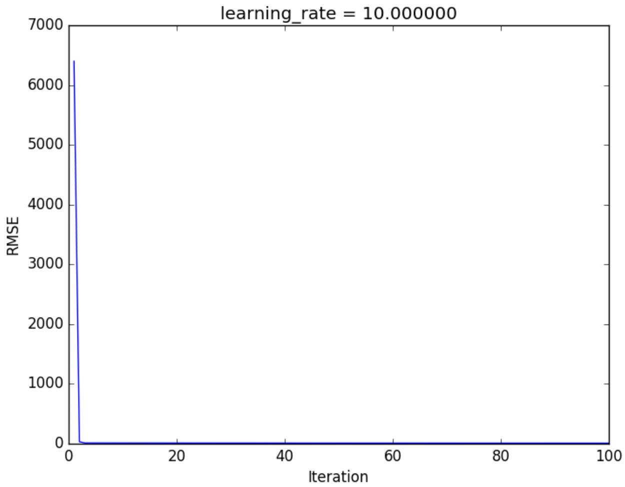
1. 請分別使用至少 **4** 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。

以下為示意圖：









以上四張圖比較 learning rate 分別為 0.01、0.1、1、10 的情況，training data 只取 9 小時內 PM2.5 的一次項，疊代次數為 100 次。在更多次疊代後，四筆資料的結果幾乎相同，所以使用更少的次數觀察初始的變化。

learning rate 為 0.01 時，一開始 RMSE 有較顯著的下降，在約 15 次疊代後曲線趨於平緩，並穩定下降。在 100 次之後 RMSE 約停在 11~12 之間，如果能給予更多的次數會得到和其他圖相似的最終結果。

learning rate 為 0.1 時，有比前圖更快速的下降。之後經過兩個明顯轉折(可能是 local minization)，後平緩下降。100 次後 RMSE 約停在 8 上下。

learning rate 為 1 時，初始的下降速度和前途相差不大，但以更紹的疊代次數達到 RMSE = 8 的數值。learning rate 為 10 時以和前途相差無幾。

較小的 learning rate 需要花費更多的疊代次數達到更小的 RMSE，較大的 learning rate 因為 gradient 移動的幅度大，則可能會遺漏可能的答案。

2. 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）。

在 kaggle 上的分數：

所有 feature 的一次項 + bias 項為	8.25091
只有 PM2.5 的一次項 + bias 項為	8.72254

相較於只討論過去的 PM2.5，考慮更多參數更有可能模擬真實情況像：PM10、SO_4、降雨量等... 很多參數直接或間接影響了 PM2.5 的數值，但要面臨無關連的資料混入的風險，如果有充足的背景知識則可以做更是當的篩選；再者，考慮所有資料則有 162 個一次項，資料量是否足夠以避免 over fit 是另一個考慮點。已結果來說，使用所有的資料相較於只使用 PM2.5 有較佳的表現，但不能因此認為它是最好的模型。

3. 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。

抽出 240 筆資料作 RMSEloss

λ = 0 ，即不使用 regulization 時 training RMSE 為 4109 ， kaggle 的分數為 6.68026。weight 的 L2 norm 為 102.54

λ = 0 .01 時 training RMSE 為 21680 ， kaggle 未上傳，但 test data 從 200 多至-200 多，誤差必定極大。weight 的 L2 norm 為 2277.34

λ = 0 .0001 時 training RMSE 為 4846 ，kaggle 的分數約 53，預測結果仍大量失真。weight 的 L2 norm 為 1086.68

λ = 0 .000001 時 training RMSE 為 4096 ，kaggle 的分數約 6.70355，和原本的預測接近。weight 的 L2 norm 為 101.25

λ = 0 .00000001 時 training RMSE 為 4109 ，kaggle 的分數約 6.67480，和未做 regulization 分數已經相差不多。weight 的 L2 norm 為 102.53

regulization 的用意是使曲線更平滑一些降低多項的影響，但參數過大時會使曲線傾向於低次，而無法擬合參考點，如 λ = 0 .1 時的結果。而 λ 調至過小，如 λ = 0 .00000001，

則看不出使用 regulization 的效果。只能多次嘗試，設法找出最佳的 λ 值。以上述結果看來 weight 的幾和平均和 RMSE 略成正相關，有高 weight-L2norm 的 data 有高機率有高 RMSE-loss。

4.

4.a

利用 SSE 對 w 微分等於 0 求得最佳的 w^*

定義：$\widehat{r_n}$ $< r_1, r_2 ... r_N >$ 的對角矩陣 (只有對角線上有值)

定義：$\widehat{x_n}$ $< x_1, x_2 ... x_N >$ 每個 x 項是一個 vector

$$SSE = E_D(w) = \frac{1}{2}\sum_{n-1}^{N} r_n(t_n - w)^2$$

$$= \frac{1}{2}(\widehat{x_n}w - t_n)^T\widehat{r_n}(\widehat{x_n}w - t_n)$$

$$= \frac{1}{2}(w^T\widehat{x_n}^T - t_n^T)\widehat{r_n}(\widehat{x_n}w - t_n)$$

$$= \frac{1}{2}(w^T\widehat{x_n}^T\widehat{r_n} - t_n^T\widehat{r_n})(\widehat{x_n}w - t_n)$$

$$= \frac{1}{2}w^T\widehat{x_n}^T\widehat{r_n}\widehat{x_n}w - w^T\widehat{x_n}^T\widehat{r_n}t_n - t_n^T\widehat{r_n}\widehat{x_n}w - t_n^T\widehat{r_n}t_n$$

對 w 微分

$$\frac{\partial E_D(w)}{\partial w} = \frac{1}{2}(2\widehat{x_n}^T\widehat{r_n}\widehat{x_n}w - 2t_n^T\widehat{r_n}\widehat{x_n})$$

$$= \widehat{x_n}^T\widehat{r_n}\widehat{x_n}w - t_n^T\widehat{r_n}\widehat{x_n}$$

$$w^* = (\widehat{x_n}^T\widehat{r_n}\widehat{x_n})^{-1}t_n^T\widehat{r_n}\widehat{x_n}$$

4.b

先進行轉置，以利計算過程

$$t_n = \begin{pmatrix} 0 \\ 10 \\ 5 \end{pmatrix}$$

$$\widehat{x_n} = \begin{pmatrix} 2 & 5 \\ 5 & 3 \\ 1 & 6 \end{pmatrix}$$

$$\widehat{r_n} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

使用 Matlab 輔助運算

$$w^* = \begin{pmatrix} 0.0560 & -0.0472 \\ -0.0472 & 0.0476 \end{pmatrix} t_n^T \widehat{r_n} \widehat{x_n}$$

$$= \begin{pmatrix} 0.0560 & -0.0472 \\ -0.0472 & 0.0476 \end{pmatrix} \widehat{x_n}^T \widehat{r_n} t_n$$

$$= \begin{pmatrix} 2.2828 \\ -1.1359 \end{pmatrix}$$

回復轉置狀態

$$w^* = \begin{pmatrix} 2.2828 & -1.1359 \end{pmatrix}$$

5.

將新的資料放入函式

$$y_{new}(x_n, w) = w_0 + \sum_{d=1}^{D} w_d x_{nd} + \sum_{d=1}^{D} w_d \varepsilon_{nd}$$

$$= y(x_n, w) + \sum_{d=1}^{D} w_d \varepsilon_{nd}$$

$$E_{D,new}(w) = \frac{1}{2} \sum_{n=1}^{N} (y_{new}(x_n, w) - t_n)^2$$

$$= \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) + \sum_{d=1}^{D} w_d \varepsilon_{nd} - t_n)^2$$

$$E[E_{D,new}(w)] = \frac{1}{2} \sum_{n=1}^{N} E[(y(x_n, w) + \sum_{d=1}^{D} w_d \varepsilon_{nd} - t_n)^2]$$

$$E[E_{D,new}(w)] = \frac{1}{2} \sum_{n=1}^{N} E[(y(x_n, w) - t_n)^2 - 2(y(x_n, w) - t_n) \sum_{d=1}^{D} w_d \varepsilon_{nd} + (\sum_{d=1}^{D} w_d \varepsilon_{nd})^2]$$

$$= \frac{1}{2}\sum_{n=1}^{N} E[(y(x_n,w)-t_n)^2] - 2E[(y(x_n,w)-t_n)\sum_{d=1}^{D} w_d\varepsilon_{nd}] + E[(\sum_{d=1}^{D} w_d\varepsilon_{nd})^2]$$

由於 $E[\varepsilon_n]$ 為 0 ，中間項可刪除

$$= \frac{1}{2}\sum_{n=1}^{N} E[(y(x_n,w)-t_n)^2] + E[(\sum_{d=1}^{D} w_d\varepsilon_{nd})^2]$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N} E[(\sum_{d=1}^{D} w_d\varepsilon_{nd})^2]$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N} E[(\sum_{d=1}^{D} w_d\varepsilon_{nd})(\sum_{k=1}^{D} w_k\varepsilon_{nk})]$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N} E[\sum_{d=1}^{D}\sum_{k=1}^{D} w_d\varepsilon_{nd}w_k\varepsilon_{nk}]$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{k=1}^{D} w_d w_k E[\varepsilon_{nd}\varepsilon_{nk}]$$

$$= E_D(w) + \frac{1}{2}\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{k=1}^{D} w_d w_k \delta_{dk}\sigma^2$$

$$= E_D(w) + \frac{\sigma^2}{2}\sum_{n=1}^{N}\sum_{d=1}^{D} w_d^2$$

$$= E_D(w) + \frac{N\sigma^2}{2}\sum_{d=1}^{D} w_d^2$$

獲得除了 bias 外 L_2-norm 的類型

6.

矩陣可逆時

$$\frac{d}{dA_{ij}} ln(det(A)) = \frac{adj(A)}{det(A)} = A_{ji}^{-1}$$

搭配連鎖率

$$\frac{d}{d\alpha} ln(det(A)) = \frac{adj(A)}{det(A)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{dA_{ij}(\alpha)}{d\alpha}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ji}^{-1}(\frac{d}{d\alpha}A)_{ij}$$

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