Homework 1 Report - PM2.5 Prediction

資工四 B04902131 黃郁凱

October 12, 2018

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。
   I use five learning rate $[0.1, 0.05, 0.01, 0.005, 0.001]$ to test its convergence efficiency and effectiveness. As we can see in the figure, too large or too small learning rate is not helpful for the training. When learning rate is too high, it is not recommended because the gradient is only a good local approximation of the loss function. When the step size is too small, it goes to slowly to jump out a worse local minima. Carefully observed from the loss values, I find it more suitable to set learning rate as 0.005 as far as efficiency and effectiveness are concerned.

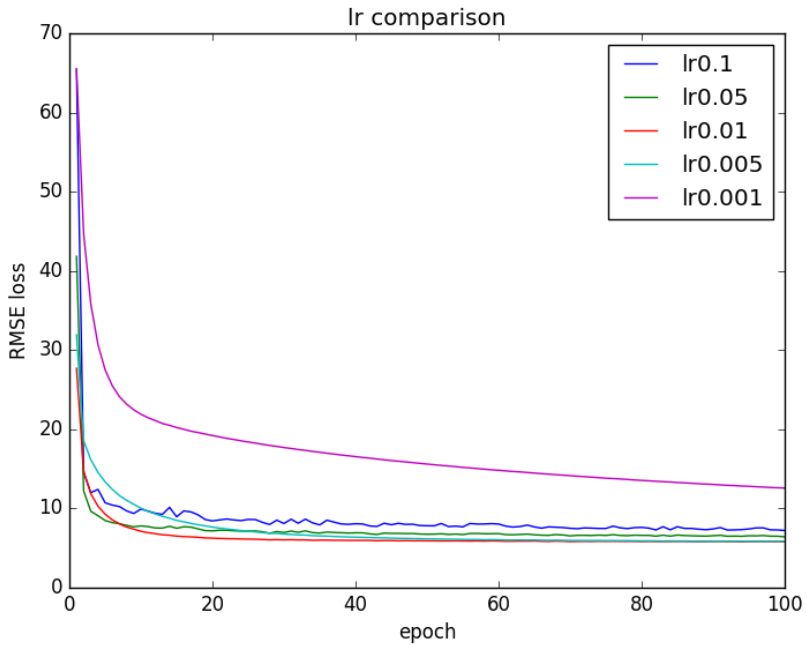

Figure 1: The plot figure of RMSE loss versus training epoch under different learning rates.

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）。

When I use all features to train, I get 10.06 on public score. But, I get 7.84 using only PM2.5 feature. The huge difference 2.22 comes from some features, like **WIND_DIREC** and **WD_HR**. These features barely relate to **PM2.5** but have large variational values such that prediction of linear regression deteriorates. Linear regression is sensitive to outlier or biased data, so those features influence the performance.

To make sure that the feature indeed influece the results, I dump the parameters of linear regression and find that the average parameter values of **WIND_DIREC** is ten times more than **PM2.5**. The results tell that before training, make sure all the feature in training data is helpful.

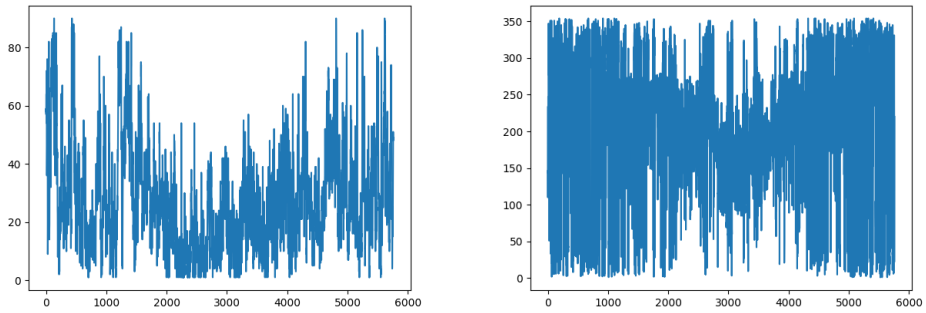

Figure 2: The plots of features: PM2.5(left), WIND_DIREC(right) of values versus time. I can see that the values of WIND_DIREC are large, change rapidly, and has no direct relation compared with PM2.5.

3. 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一致），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。
   The results tell that the less regularization is better. The conclusion makes sense since the purpose of regularization is to weaken the power of the model by imposing a penalty on the complexity of the function. Linear regression is such a simple model that when forcing it to be simplier, I get a deteriorated model instead. Also, note that the usage of regularization is to prevent overfitting, in the homework case, it is unnecessary to do this since I suffer little overfitting.

λ	training	testing	
		public	private
10^{-4}	7.15	8.85	8.29
10^{-5}	6.09	7.04	7.01
10^{-6}	5.69	6.28	6.59
10^{-7}	5.64	6.12	6.55
0.0	5.65	6.09	6.56

4. (a) Given t_n is the data point of the data set $\mathcal{D} = \{t_1, ..., t_N\}$. Each data point t_n is associated with a weighting factor $r_n > 0$. The sum-of-squares error function becomes:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n(t_n - \mathbf{w^T x}_n)^2$$

Find the solution $\mathbf{w}^*$ that minimizes the error function.

let $R = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n \end{bmatrix}$, $T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$, $X = \begin{bmatrix} - - x_1 - - \\ - - x_2 - - \\ \vdots \\ - - x_n - - \end{bmatrix}$.

Now the error function becomes $E_D(\mathbf{W}) = \frac{1}{2}(XW - T)^T R(XW - T)$. To find minimum value, let

$$\begin{aligned} \nabla_{\mathbf{W}} E_D(\mathbf{W}) &= 0 \\ &= \nabla_{\mathbf{W}} \frac{1}{2}(XW - T)^T R(XW - T) \\ &= \frac{\partial \frac{1}{2}(W^T X^T RXW - W^T X^T RT - T^T RXW + T^T RT)}{\partial \mathbf{W}} \\ &= \frac{1}{2}(2X^T RXW - X^T RT - X^T RT) \\ &= X^T RXW - X^T RT \end{aligned}$$

$$\begin{aligned} \Rightarrow & X^T RXW - X^T RT = 0 \\ & W = (X^T RX)^{-1} X^T RT \end{aligned}$$

When $w^* = (X^T RX)^{-1} X^T RT$, it minimizes the error function.

(b) Following the previous problem(4-a), if

$$\mathbf{t} = [t_1 t_2 t_3] = \begin{bmatrix} 0 & 10 & 5 \end{bmatrix}, \mathbf{X} = [\mathbf{x_1 x_2 x_3}] = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}$$

$$r_1 = 2, r_2 = 1, r_3 = 3$$

Find the solution $\mathbf{w}^*$.

$$\begin{aligned}
w^* &= (X^T R X)^{-1} X^T R T \\
&= (\begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix})^{-1} \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
&= (\begin{bmatrix} 4 & 5 & 15 \\ 6 & 1 & 18 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix})^{-1} \begin{bmatrix} 4 & 5 & 15 \\ 6 & 1 & 18 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} \\
&= (\begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix})^{-1} \begin{bmatrix} 125 \\ 100 \end{bmatrix} \\
&\approx \begin{bmatrix} 2.28 \\ -1.14 \end{bmatrix}
\end{aligned}$$

5. 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:

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

where t_n is the data point of the data set $\mathcal{D} = \{t_1, ..., t_N\}$

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.

$$
\begin{aligned}
E'(w) &= \frac{1}{2}\sum_{n=1}^{N}(y(x_n+\epsilon_n, w) - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D} w_i(x_i+\epsilon_{ni}) - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} w_i\epsilon_{ni} - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) + \sum_{i=1}^{D} w_i\epsilon_{ni} - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n)\sum_{i=1}^{D} w_i\epsilon_{ni} + \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \epsilon_{ni}\epsilon_{nj}
\end{aligned}
$$

Take average over the noise distribution and get

$$
\begin{aligned}
\mathbb{E}(E'(w)) &= \mathbb{E}\left(\frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2 + 2(y(x_n, w) - t_n)\sum_{i=1}^{D} w_i\epsilon_{ni} + \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \epsilon_{ni}\epsilon_{nj}\right) \\
&= \mathbb{E}\left(\frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2\right) + 2(y(x_n, w) - t_n)\sum_{i=1}^{D} w_i\mathbb{E}(\epsilon_{ni}) + \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \mathbb{E}(\epsilon_{ni}\epsilon_{nj}) \\
&= \mathbb{E}\left(\frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2\right) + 0 + \sigma^2\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \\
&= \mathbb{E}(E(w)) + C\|w\|_2
\end{aligned}
$$

, where C is a constant.

6. $\mathbf{A} \in \mathbb{R}^{n\times n}, \alpha$ 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.

$$
\begin{aligned}
\frac{dln|A|}{d\alpha} &= \frac{\mathrm{d}ln(det(A))}{\mathrm{d}\alpha} \\
&= \frac{1}{det(A)} \frac{\mathrm{d}det(A)}{\mathrm{d}\alpha} \\
&= \frac{1}{det(A)} \sum_{i=1}^{n} \sum_{j=1}^{n} Adj(A)_{ji} (\frac{\mathrm{d}A}{\mathrm{d}\alpha})_{ij} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ji}^{-1} (\frac{\mathrm{d}A}{\mathrm{d}\alpha})_{ij} \\
&= Tr(A^{-1} \frac{\mathrm{d}A}{\mathrm{d}\alpha})
\end{aligned}
$$