作業一報告 - PM2.5 預測

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。

   我使用五種學習率 $[0.1, 0.05, 0.01, 0.005, 0.001]$ 來測試訓練的收斂效率跟效能。從圖片當中我們可以看到,太大或太小的學習率效果都不會很好。當學習率過大,效率高但效能不穩定,容易參數更新過多,導致跳到更糟的損失曲面位置,因為梯度只是一個局部的方向走勢,不代表全局的走勢。當學習率過小,參數更新緩慢、力道不夠,容易卡在比較糟的局部最低值或是平坦的損失曲面,使得效率與效能都不好。我發現當使用學習率 0.005 的時候效率與效能都達到最好,也就是更新速度夠快,並收斂到不錯的局部最低值。

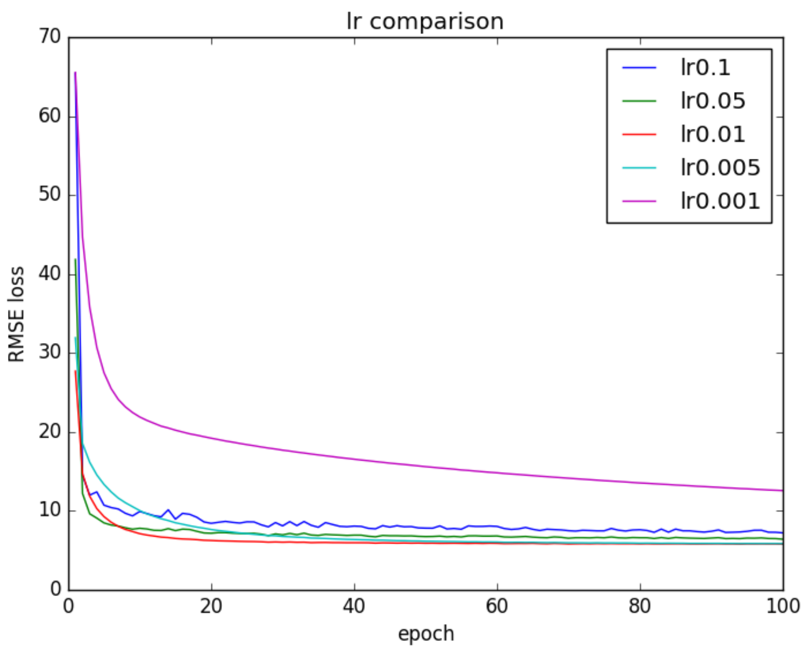

Figure 1: 不同學習率下的 RMSE 損失值相對於訓練 epoch 數

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)。

當我用所有的特徵 (總共十八項) 下去訓練，得到比較糟分數；但只用 **PM2.5** 這項特徵下去訓練就進步了 2 到 3 的差距。這樣大的差距來自於某些特徵如 **WIND_DIREC** 和 **WD_HR**，他們的數值大且有著巨大變化，並且和 **PM2.5** 的曲線走勢看不出相關性。因為線性回歸容易被大偏差的值所影響，那些可能有巨幅變化的特徵並不適合。

特徵	公開分數	最終分數
全用	10.05	11.67
用 PM2.5	7.84	8.21

Table 1: 公開與最終分數在取用不同特徵來訓練的比較，單位是 RMSE 的數值損失。

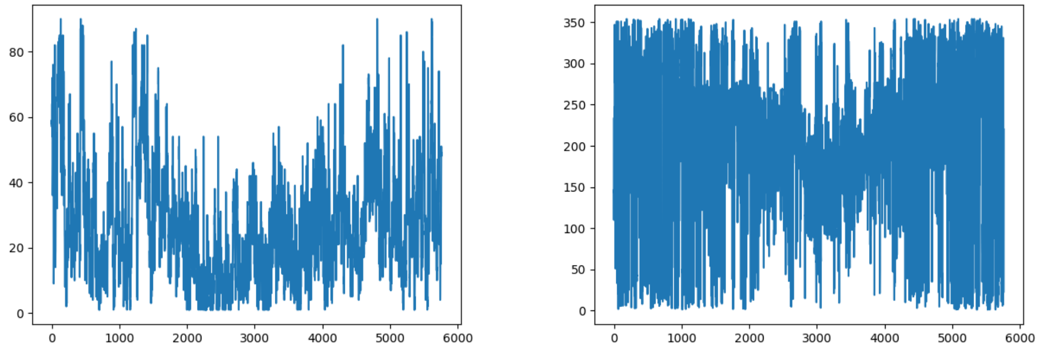

Figure 2: 不同特徵的值對時間作圖: PM2.5(左), WIND_DIREC(右)。可以明顯看出，WIND_DIREC 這項特徵的值隨時間的變化與 PM2.5 無關，並且它的值偏大且變化劇烈，容易影響線性回歸模型的預測準確。

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。

   從實驗數據可以看出，當加 regularization 對於抑制 overfit 有幫助，加越多效果越好。然而，不管公開或私人的分數都會一起變差，加越多 regularization 表現就會越差。這結果合理，因為 regularization 目的是抑制模型過強，減緩 overfit 的情形發生，但是這次使用的線性回歸模型簡單，並不用太擔心模型過強造成測試時變差很多。

λ 比重	訓練分數	測試分數	
		公開	最終
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

Table 2: 分數與 regularization 比重的比較，單位是 RMSE 的數值損失。

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.

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

Take average over the noise distribution and get

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

, where C is a constant.

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

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