

Machine Learning HW1: PM2.5 Prediction

學號: R06922030 系級: 資工碩一 姓名: 傅敏桓

請實做以下兩種不同 feature 的模型，回答第 1~3 題：

- (1) 抽全部 9 小時內的污染源 feature 的一次項(加 bias)
- (2) 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註：

- a. NR 請皆設為 0，其他的數值不要做任何更動
- b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等)都是可以用的

針對以下(1)至(3)問題，我取每個月最末端的 48 筆數據做為 validation set，模型參數初始值均由亂數決定（偏移項初始值為 1），透過 adagrad 訓練模型，進行至多 50000 次的 iteration 直到誤差在測試集上小於某個值($1e-6$)為止。

1. (2%)記錄誤差值 (RMSE) (根據 kaggle public+private 分數)，討論兩種 feature 的影響

這邊用 (a) 代表取全部 9 小時內的污染源一次項的模型，(b) 代表單抽 pm2.5 的一次項的模型。

- (a) 在 kaggle 上分別得到 public=8.17254 和 private=6.09744 的分數 (RMSE=7.21003)
- (b) 在 kaggle 上分別得到 public=7.41858 和 private=5.58012 的分數 (RMSE=6.56403)

另外，(a) 在 validation set 之 RMSE 為 6.826218，(b) 則為 7.032354。可以發現 (b) 的表現普遍都比 (a) 來得好。有較多參數的模型 (a) 是比 (b) 複雜許多，表現比較差的原因除了可能是 (a) 還沒有被訓練到最好外，我還考慮了以下兩種可能性：

- i. 我們的資料可能沒辦法單用線性模型 fit 得太好，參數過多就有 overfitting 的問題
- ii. 有一些觀測數據可能和 PM2.5 幾乎不相關，加入模型容易造成預測偏差

我認為 (b) 也能 work 的原因是 PM2.5 的值似乎不容易陡升陡降，我們若掌握到前幾個小時的觀測值，或者再考慮一些比較關鍵性的 feature，應該就足夠猜出接下來觀測值的走向。

2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

這邊用 (c) 代表取全部 5 小時內的污染源一次項的模型，(d) 代表單抽 pm2.5 的一次項的模型。

- (c) 在 kaggle 上分別得到 public=7.59881 和 private=5.31483 的分數 (RMSE=6.55703)
- (d) 在 kaggle 上分別得到 public=7.56623 和 private=5.75828 的分數 (RMSE=6.72330)

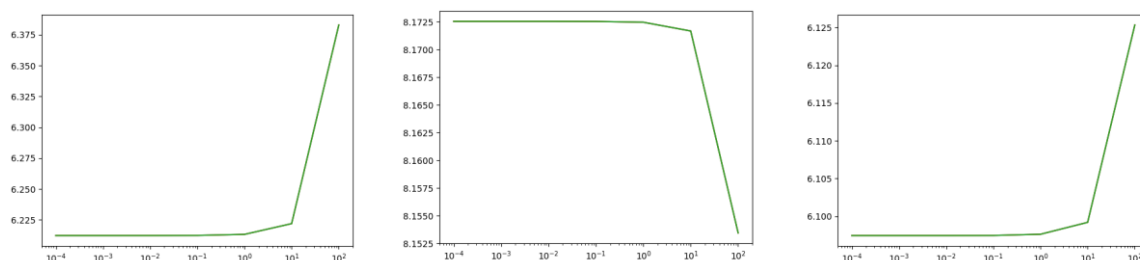
觀察模型 (c) 與上題 (a) 之差異，我推測分數進步的原因是參數變少，減輕了模型 overfitting 的現象，或者我們的訓練資料不夠多導致參數多的模型訓練的不夠好。而模型 (d) 的表現會略差於 (b)，我推測原因是 (d) 的模型參數太少，過度簡化的模型能力不足所致；因 (b) (d) 這兩個模型在訓練的時候都是停在參數更新量太小，不太可能是有哪一邊模型訓練得不好造成的。

從這兩題的結果我觀察到參數選擇所帶來的差異，參數多的模型雖然理論上預測能力比較強，但有可能過度貼合訓練資料，反而在測試資料上表現不佳；一方面參數過少的模型則能力不足以預測太複雜的資料。（推論畢竟是以 kaggle 分數為基礎，只能討論在這次的作業上各模型的表現，而不代表哪種模型真的比較好）

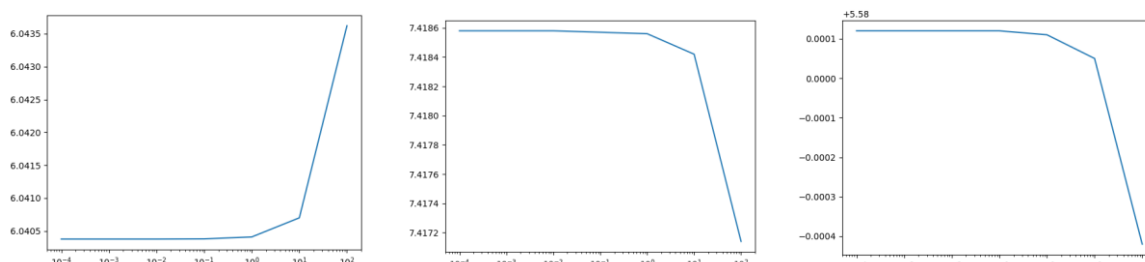
3. (1%) Regularization on all the weight with $\lambda=0.1, 0.01, 0.001, 0.0001$, 並作圖

模型設定與上述(a) (b) 相同 (只有 0.1 以下實在看不出差別, 加入 $\lambda=1.0, 10.0, 100.0$)。

以下是取全部觀測值(a)分別在 train, public 和 private test data 的變化



以下是只取 PM2.5 觀測值(b)分別在 train, public 和 private test data 的變化



可以觀察到隨 λ 值增加, 在訓練集上的誤差些微變大, 而在測試集上的些微變小 (把線畫在一起會看不出變化所以分開來畫, 實際上分數差距大約最多 0.02)。雖然可以預期越平滑的 (λ 越大) 模型越有機會在測試集上表現較好, 我認為在這次的線性模型中, 常規化的比重對於整體結果仍然影響不大。

4. (1%) 在線性回歸問題中, 假設有 N 筆訓練資料, 每筆訓練資料的特徵 (feature) 為一向量 x^n , 其標註 (label) 為一純量 y^n , 模型參數為一向量 w (此處忽略偏權值 b), 則線性回歸的損失函數 (loss function) 為 $\sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1 \ x^2 \ \dots \ x^N]^T$ 表示, 所有訓練資料的標註以向量 $y = [y^1 \ y^2 \ \dots \ y^N]^T$ 表示, 請問如何以 X 和 y 表示可以最小化損失函數的向量 w ? 請寫下算式並選出正確答案。(其中 $X^T X$ 為 invertible)

- (a) $(X^T X) X^T y$ (b) $(X^T X)^{-1} X^T y$ (c) $(X^T X)^{-1} X^T y$ (d) $(X^T X)^{-2} X^T y$

給定訓練資料的特徵及對應之標註, 損失函數可以定義為 $L(w) = \sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。我們的目標是找出一組模型參數 w^* 使得損失函數最小, 即 $w^* = \arg \min_w L(w)$ 。藉由矩陣乘法的運算規則, 我們可以將損失函數改寫為向量內積的形式 $(X \cdot w - y)^T (X \cdot w - y)$, 然後將 $L(w)$ 對 w 微分:

$$\begin{aligned} \frac{\partial L(w)}{\partial w} &= \frac{\partial}{\partial w} (Xw - y)^T (Xw - y) \\ &= \frac{\partial}{\partial w} ((Xw)^T - y^T) (Xw - y) \\ &= \frac{\partial}{\partial w} (w^T X^T Xw - w^T X^T y - y^T Xw + y^T y) \\ &= ((w^T X^T X)^T - (y^T X)^T) \\ &= (X^T Xw - X^T y) \end{aligned}$$

(根據矩陣導數之定義, 對任意向量 x 及 $u(x)$, 若 A 非 x 的函數, 則有 $\frac{\partial x}{\partial x} = I$, $\frac{\partial Au}{\partial x} = \frac{\partial u}{\partial x} A^T$)

取以上微分為 0 得到 $w^* = (X^T X)^{-1} X^T y$, 答案為 (c) 選項。