# 機器學習作業一 report

學號：B05901040　系級：電機三　姓名：蔡松達

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 當作一次項(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

c. 第 1-3 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好，kaggle 死線之後便可以無限上傳。

e. 根據助教時間的公式表示，(1) 代表 p = 9x18+1 而(2) 代表 p = 9*1+1

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

| RMSE | Training | public | private | Testing average |
|---|---|---|---|---|
| All 9hr | 5.52803 | 5.65432 | 7.24981 | 6.45207 |
| PM2.5 9hr | 5.98009 | 5.94811 | 7.45901 | 6.70356 |

Training 此處使用 adagrad，learning rate=1，epochs=10000。從此狀況可看出使用全部的汙染源能夠帶來明顯較好的 training 及 testing 結果，不僅是因為有比較多的參考資訊，更是因為許多汙染源直接與 PM2.5 有密切關聯，包含 $NO_2$、$SO_2$、CO，雨量也有所關聯（雖然大部分數值為 0，但有降雨時 PM2.5 值將顯著降低），因此使用所有汙染源資訊能夠以更多的視角來預測，相對於只使用 PM2.5，在 public 及 private 都有較好結果。

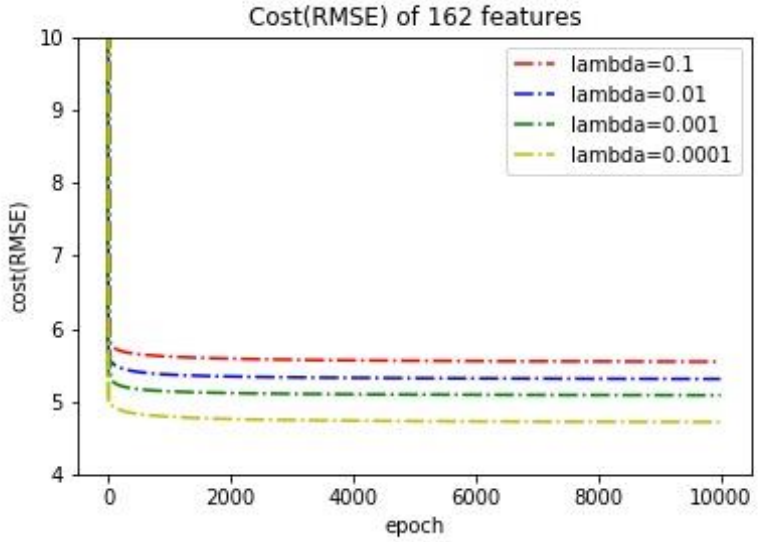
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

| RMSE | Training | public | private | Testing average |
|---|---|---|---|---|
| All 5hr | 5.12239 | 5.97626 | 7.20289 | 6.58958 |
| PM2.5 5hr | 5.92813 | 6.31862 | 7.39042 | 6.85452 |

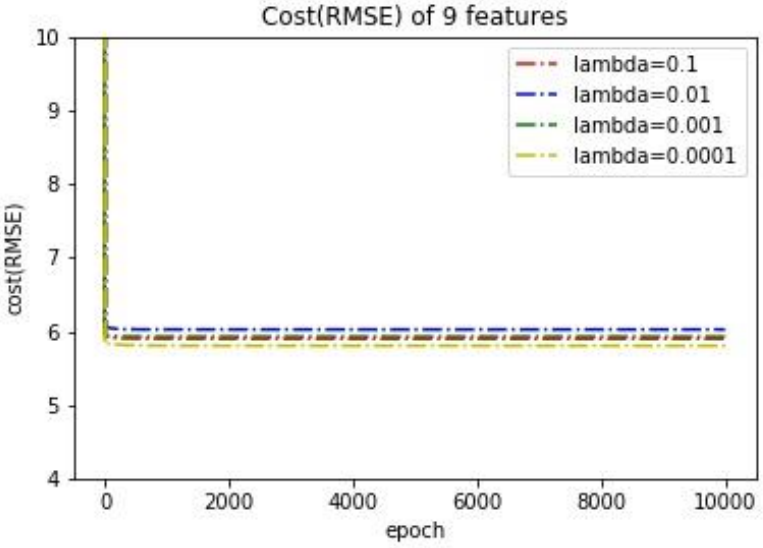
Training 此處亦使用 adagrad，learning rate=1，epochs=10000。之所以在 training 上可以得到較第一題好的結果可能是因為其 feature 使用較少，但在 testing 則得到相對於取九小時較差的結果，主因是因為九小時的 feature 終究能提供較多資訊，例如九小時內可能有降雨，但若只取五小時便可能會失去此資訊；次因也可能是因為有些 overfit，因此即便 training 變好，testing 仍舊變糟不少。亦可從五小時全取與九小時只取 PM2.5 做比較，依舊是全取汙染源的表現較佳，可見單純使用 PM2.5 進行 training 無法有太好的效果。而如同第一題的趨勢，同樣只取五小時的狀況下，全取汙染源較只取 PM2.5 為佳。

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

(1) 抽全部 9 小時內的污染源 feature 當作一次項(加 bias)：Training 此處使用 adagrad，learning rate=1，epochs=10000。圖中顯示使用的 λ 越小能夠得到較好的 training 結果，可能的原因 λ 太大時可能會太過主導導致 w 趨近於 0，無法 train 出較符合 training data 的 model；而收斂的速度基本上都是在 100 epochs 就不會再變的更好。




(2) 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias) ：Training 此處使用 adagrad，learning rate=1，epochs=10000。圖中無法明顯看出 λ 與收斂之後的 cost 大小之間的關聯性，可能是因為採用的 feature 量相當少，收斂的狀況較不會受到其他因素影響，λ 可能造成的效應比起全取汙染源不顯著，cost 不太容易繼續往下走；而收斂的速度基本上都是在 100 epochs 就不會再變的更好。（此處有進行 training data 的 shuffle，能夠主導 cost 收斂的結果也可能有此因素存在）



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 n=1Nyn-xnw2 。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 ？請選出正確答案。(其中 $X^TX$ 為 invertible)

a. $(X^TX)X^Ty$

b. $(X^TX)yX^T$

c. $(X^TX)^{-1}X^Ty$

d. $(X^TX)^{-1}yX^T$

答案為 C