Machine learning HW1 Report R05942056 時丕澔

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：至此報告撰寫為止，我嘗試使用的輸入特徵有 PM2.5、PM10 和風的角度、大小。

處理方式為：

(1). 將這些對應的資料從主資料表中抽取出來，分別存成大矩陣(以練習提供的資料 train.csv 為例，分別為 240 x 24 的矩陣)

(2). 為了增加學習 data 量，將所有可能的「第十個小時的 PM2.5 資料」抽取出來當作 training data 的輸出 $Y = [y_1, y_2 \ldots . y_{5751}]^T$（5751 = 240*24-9），再將對應的「前九個小時」的資料抽取出來當作 training data 的輸入 $X = [X_1, \ldots X_N]$（每個 X_n 皆為長度為 5751 的行向量。）

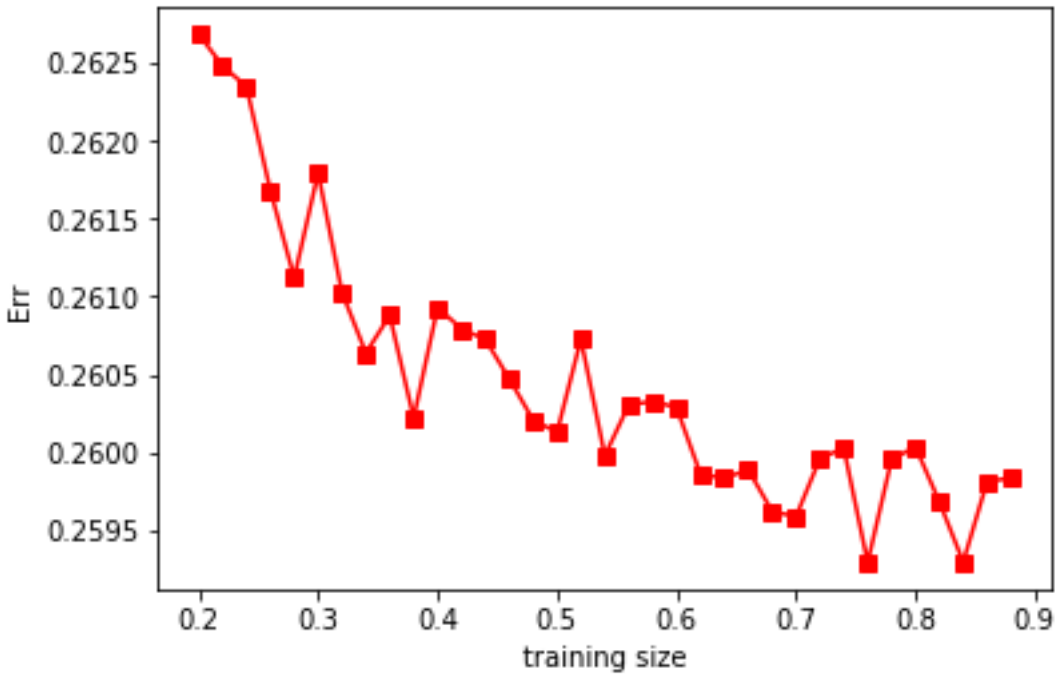
(3). 對有關風的資料（風的大小、方向）作處理，採取簡化的 model：假設 PM2.5 值和前九個小時的「投影至某個方向的風的大小」有線性關係，寫成式子就是每個小時都會有一個代表風的值= (風速)*sin(風角度*pi/180+phs)，其中 phs 是一個可以調的參數，代表要取哪個方向的風向。

對 PM2.5、PM10 這些指標值的資料，則不做這些轉換。

(4). 對所有 feature 作 regularization。

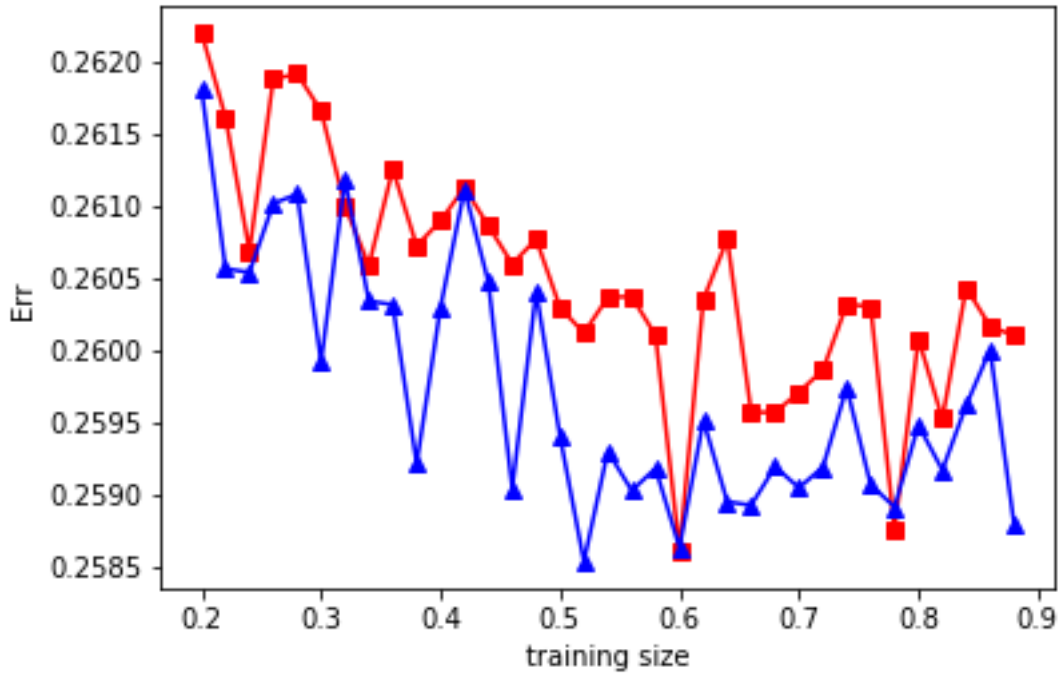
2.請作圖比較不同訓練資料量對於 PM2.5 預測準確率的影響

答：下圖為使用所有 training data 作隨機取樣，testing data 取其中所有 data 的 10%，training data 分別取所有 data 的 20~90%（testing 和 training data 兩者不重複），分別作 5000 次實驗（計算過程有使用 close form 的求解方法以加速運算）後平均的結果。可以發現雖然越多的資料、預測效果有越好、誤差越小的趨勢。

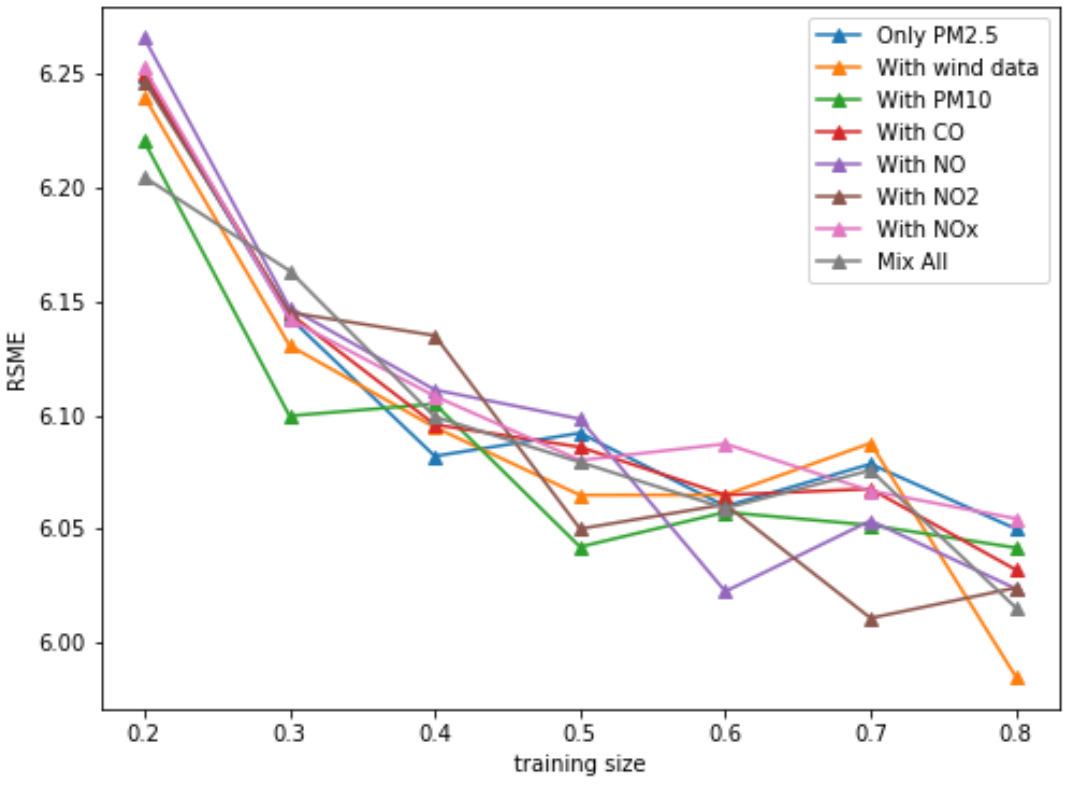


3. 請比較不同複雜度的模型對於 PM2.5 預測準確率的影響

答：下圖為使用的 feature 多寡不同的兩種情況，計算過程同 2.。差別在於下圖中紅色線為使用風場資料預測的結果，藍色為不使用風場資料預測的結果。可以看出即使前者的模型複雜度高於後者，預測結果仍不一定較佳，這和使用模型的優劣有密切關聯。



下圖是使用各種不同 feature 實驗後的比較結果，可以看出最簡單的只使用 PM2.5 的模型普遍會比有添加其他 feature 的模型表現還要遜色，但使用最多 feature 的（灰色）也並非一直是最好的。



4. 請討論正規化(regularization)對於 PM2.5 預測準確率的影響
答：在 Linear regression 問題中，因為 feature 的平移與縮放（正規化的兩個部分）等價於參數的平移與縮放，顧最終等於只是對 loss function 的參數域在每個維度上作 scaling，loss function 本身的值並不會改變，也就是說最小值也不會改變；故如果在正規化前後的 Linear regression 問題中參數皆是取到使 loss function 最小的（這也是 Linear regression 可以輕易做到的），他們對應到的 model 也會是同一個。所以理論上對任何 testing 輸入，testing 輸出不會有任何差別（差別只在於收斂過程會有顯著改變）。

5. 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)為一存量 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$\sum_{n=1}^{N}(y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 X = $[x^1\ x^2 \dots x^N]$ 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，請以 X 和 y 表示可以最小化損失函數的向量 w 。
答：
這只是簡單的線性代數。
$w = (X^TX)^{-1}X^Ty$