學號：R05921087 系級： 電機碩一 姓名：鍾智堅

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

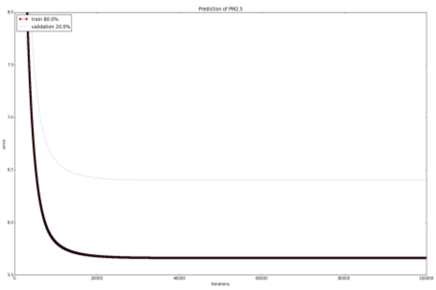
答：

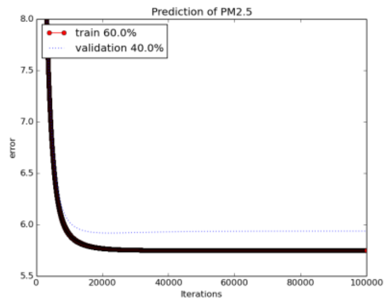
參考了之前學長姐寫過的作業和其他研究論文，第一次我選擇前九小時的 PM10 和 PM2.5 為 feature，但是發現模型是有偏向非線性的成長。

因此，我調整最好的模型，是抽取第 2 小時到第 9 小時的 PM2.5 作為 feature，即 feature 有 8 個。

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：





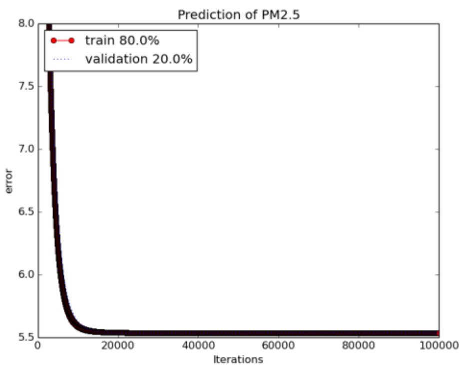
左圖為訓練資料量是原有資料量的 80%，右圖的訓練資料量為 60%。左圖的 PM2.5 的 training error 為 5.664、validation error 為 6.403；右圖 PM2.5 的 training error 為 5.749、validation error 為 5.935。

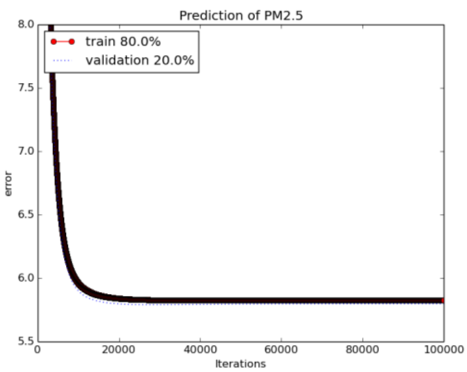
當訓練資料量比較高，自然有足夠量不同的樣本以形成更好的模型，但是 validation 資料量比較少，因此誤差更大。反過來，訓練資料量低，因此誤差高，但是 validation 資料量比較多，因此把平均誤差壓低了。

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：

模型越複雜，那麼其表現是越靠近非線性。在利用 linear regression 中，模型越複雜，準確率也開始降低。





左圖的 feature 為前 9 小時的溫度、CO、PM10、PM2.5、雨量以及 SO2，右圖的 feature 為前 9 小時的 PM2.5。左圖的 kaggle score 為 5.83549，右圖為 5.67818。由此可見，比較少 feature 的模型（即複雜度越低），自然呈現靠近線性的預測結果。

4. 請討論正規化(regularization)對於 PM2.5 預測準確率的影響

答：

Regularization 即也是調整 learning rate，可以減低 overfit 對預測準確率的負面影響。如果 learning rate 調整小、database 龐大，然後無限制 iteration 的次數，那麼 PM2.5 預測準確率會提升。

如果沒有 regularization，在 overfit 的情況下，那麼在 training 過程中，training error 會變得小，但是 validation error 卻會增大，即對 PM2.5 的預測準確率會變差。

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 $X = [x^1\ x^2 \ldots x^N]$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \ldots y^N]^T$ 表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

答：