Machine Learning 2017 Spring
Homework 1 Report

學號：B03902048
系級：資工三
姓名：林義聖

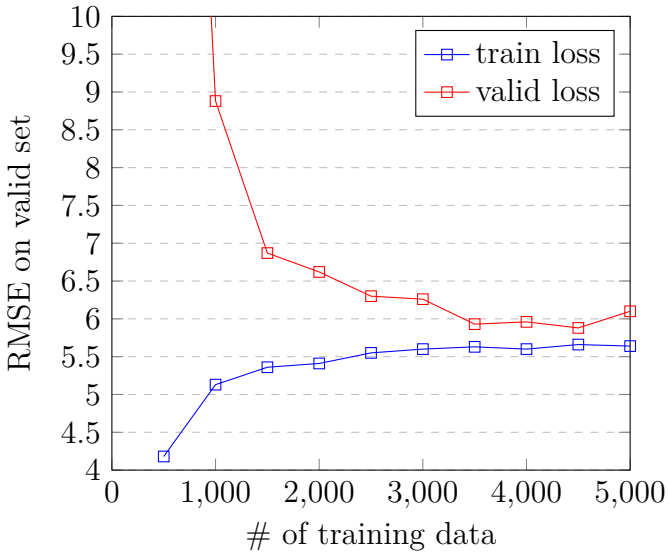
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

   答：我選擇的 feature，是根據對 PM2.5 的基本認識，先選了：PM10, PM2.5, RAINFALL, WIND_DIREC, WIND_SPEED, WD_HR, WS_SR。之後又查了相關資料，得知：CO, O3, SO2 皆與 PM2.5 相關，於是也將這些選入，用來訓練 `hw1.sh` 所執行的模型內。而後，為了進一步提升成績，我再將前述所有 features 的平方項和 PM2.5 與 O3 的乘積項加進 train_X 裡面，並用 `hw1_best.sh` 訓練出更好的模型。

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：我用隨機的方法選取 train data 和 valid data，將 valid data 的數量固定為 652 筆。而下圖呈現的 RMSE 是將每種資料量所得出的五筆結果，經過平均後的數值 (備註：valid loss 在只有 500 筆訓練資料時為 22.11，超出圖表範圍而無法顯示)。

Relation between number of training data and prediction accuracy



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：這裡使用的 features 為第一題所提到的那 10 種。

#	x^2	x^3	**entry-wise product**	**training loss**	**public-set loss**
1				5.80	5.79
2	V			5.72	5.59
3	V		PM2.5 * O3	5.68	5.53
4	V		PM2.5 * O3 PM2.5 * PM10	5.64	5.57
5	V		PM2.5 * O3 PM2.5 * SO2	5.68	5.58
6		V	PM2.5 * O3	5.68	5.57

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

   答：以下結果皆是使用前面提及的 10 種 features，且使用了一次方及二次方項。從圖表中呈現的結果來看，看得出 regularization 對於 PM2.5 預測準確率是有幫助的。以第一、二筆為一組，與第三、四筆為一組的結果比較起來，後者多了一項 entry-wise product，同時加了 regularization 之後的結果也進步一些些，可以解釋為正規化在訓練參數較多時，避免模型過度擬合，因而帶來在 public-set score 上的進步。

#	**entry-wise product**	**regularization**	**training loss**	**public-set loss**
1			5.72	5.59
2		V	5.74	5.59
3	PM2.5*O3		5.72	5.47
4	PM2.5*O3	V	5.74	5.46

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 $\sum_{n=1}^{N}(y^n - w \cdot x^n)^2$。若將所有訓練資料的特徵值以矩陣 $X = \left[x^1, x^2, \ldots, x^N\right]$ 表示，所有訓練資料的標註以向量 $y = \left[y^1, y^2, \ldots, y^N\right]^T$ 表示，請以 X 和 y 表示可以最小化損失函數的向量 w。

   答：$w = (X^T X)^{-1} X^T y$