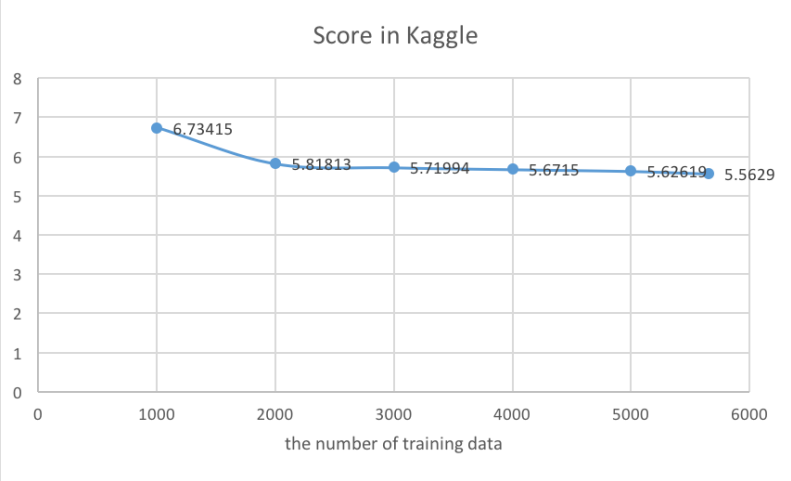
學號：B03901101 系級： 電機三 姓名：楊其昇

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：
前九個小時的所有空氣污染指標並增加前九個小時 PM10 資料的平方項以及前九個小時 PM2.5 的平方項和立方項：train_x = [x1, x2, x3…, x162, x10^2, x11^2…, x27^2, x19^3, x20^3…, x27^3]

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：



我們可以發現，當所有條件（ex: iteration 的次數，training data 的 features…等）相同時，若 training data 的數量愈多，預測的準確率會愈高（準確率愈高，Kaggle 上的分數愈低）。不過，training data 的數量愈多，準確率增加的速度會趨緩。

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：
以下所有測試結果均為在使用 AdaGrad，且 iteration = 15000, learning rate = 0.5 的情況，此外有另外做 feature normalization 以加速收斂：
(1) 增加不同種類的 feature

Feature 種類	Kaggle 上的分數
PM2.5	5.79984
PM2.5, PM10	5.86664
PM2.5, PM10, O3	5.86562
PM2.5, PM10, O3, RAINFALL	5.86206
PM2.5, PM10, O3, RAINFALL, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR	5.73302

(2) 只單純看 PM2.5 這項 feature

PM2.5 的次方項	Kaggle 上的分數
PM2.5	5.79984
PM2.5+PM2.5^2	5.74901
PM2.5+PM2.5^2+PM2.5^3	5.69177
PM2.5+PM2.5^2+PM2.5^3+ PM2.5^4	5.68565

囿於每天能上傳 Kaggle 的次數有限，因此簡單做了兩方面的比較。如果只單純增加不同的 feature，如果選到的輸入特徵不幸跟 PM2.5 的關係不大時，預測的準確率便會變得比較糟；若選擇增加 PM2.5 的次方項，準確率基本上都會獲得改善，但同時可能也會有 overfitting 的問題。因此實際上在做這個作業時，便如第一題的回答一樣，在調整參數的兩個大方向做出折衷的選擇，以達到提升準確率並同時避免產生 overfitting。

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

答：

正規化只會影響 loss function 而已，理論上會讓預測出來的模型更平順。但根據這次作業的經驗，如果選擇做正規化，為了降低做正規化所帶來的 trade-off，其他參數會變得很難調，而且預測的值和沒做正規化相比其實並沒有變得特別好，因此在這次的作業中，我後來並沒有選擇對模型做正規化。

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

答：

$$\sum_{n=1}^{N}(y^n - w \cdot x^n)^2 = (y^1 - w \cdot x^1 \ \ldots \ y^N - w \cdot x^N)\begin{pmatrix} y^1 - w \cdot x^1 \\ \vdots \\ y^N - w \cdot x^N \end{pmatrix}$$

$$= (y - Xw)^T(y - Xw)$$

$$\frac{\partial}{\partial w}(y - Xw)^T(y - Xw) = \frac{\partial}{\partial w}(y^Ty - y^TXw - w^TX^Ty + w^TX^TXw)$$

$$= X^TXw - 2X^Ty$$

嘗試使 $X^TXw - 2X^Ty = 0 \xrightarrow{yields} w = (X^TX)^{-1}X^Ty$