Machine Learning HW1

學號：R04922169 系級：資工所碩二 姓名：楊智偉

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：我試過三種方式來抽取模型所需要的特徵，其中取前 9 小時的資料再加上 bias。

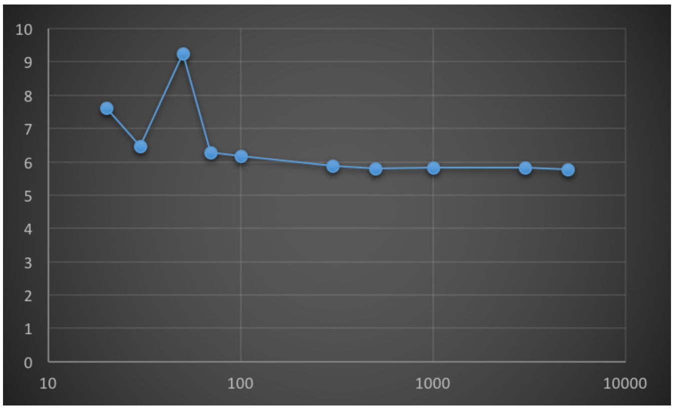
(1) 使用全部的資料來當作特徵：
18*9+1=163 維度的特徵，並將每個月的所有連續 10 小時計算進去。

(2) 使用 PM2.5 資料來當作特徵：
1*9+1=10 維度的特徵，並將每個月的所有連續 10 小時計算進去。

(3) 使用部分的 PM2.5 資料當作特徵：
1*9+1=10 維度的特徵，取出每一天 0~9 和 14~23 兩筆資料。

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：



我分別對不同的 training data 資料量計算並檢測其準確率。從上面的比較圖很明顯在剛開始準確率明顯的偏低，且非常的不穩定，然後大概在 training data 達到約 300 筆以上就會趨近到 5.7~5.8 的 MSE。然而越大的 training 量卻不一樣越準確，像是我測到 500 筆 training data 就比 1000 筆資料的結果更準確。

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：

方法	Public score
1. 使用全部的資料當特徵	6.07821
2. 使用 PM2.5 資料來當作特徵	5.79969
3. 使用部分 PM2.5 資料來當作特徵(一天抓兩筆)	5.64209
4. 將 2.的資料加入二次項特徵	5.75707

我一開始直接將所有的特徵放入模型來訓練，但是很明顯地看到結果不甚理想，推測因為使用了太複雜的參數，且其中很多參數應該與 PM2.5 的數值關係不大。後來我改由只使用 PM2.5 的資料來建立模型，就可以明顯看到準確率的提升。接著，我也嘗試過利用二次項的參數特徵來訓練，而這也取得比一次項還要好一點點的效果。而我找到最好的方法則是只使用 PM2.5 的資料，並且每一天只取兩筆資料，這個模型的結果意外的好。我推測是因為使用了太多的資料容易得到 outlier，這些 outlier 可能會造成我們 train 出來的結果不好，使用複雜低的模型則會被比較少的 outlier 所影響。

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

答：

方法	Public score
Lambda = 0	5.79969
Lambda = 0.001	5.79969
Lambda = 0.01	5.79969
Lambda = 0.1	5.79969

在這次的作業中，我覺得 regularization 的影響程度並不大。我將 random initial 的因素屏除後，針對不同 lambda 值的準確率結果做比較，卻發現他們都是一樣的。我認為這是因為 train 出來的 weight 數值都偏小，且這個 PM2.5 預測問題 fit 出來的 linear function 本身就是非常 smooth 的，所以導致此結果。

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 $X = [x^1\ x^2\ \ldots\ x^N]$ 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：

$$\text{solve } \mathrm{Xw} = \mathrm{y}, \text{ minimize } \|\mathrm{Xw} - \mathrm{y}\|^2$$

$$\Rightarrow \frac{\partial}{\partial \mathrm{w}} \|\mathrm{Xw} - \mathrm{y}\|^2 = 0 \qquad \text{(一次微分等於零)}$$

$$\Rightarrow \left[\frac{\partial}{\partial \mathrm{w}}(\mathrm{Xw} - \mathrm{y})\right][2(\mathrm{Xw} - \mathrm{y})] = 0 \qquad \text{(微分連鎖律)}$$

$$\Rightarrow \mathrm{X}^{\mathrm{T}}[2(\mathrm{Xw} - \mathrm{y})] = 0 \qquad \text{(微分連鎖律)}$$

$$\Rightarrow \mathrm{X}^{\mathrm{T}}\mathrm{Xw} = \mathrm{Xy} \ \Rightarrow \mathrm{w} = (\mathrm{W}^{\mathrm{T}}\mathrm{W})^{-1}\mathrm{W}^{\mathrm{T}}\mathrm{y}$$