學號：r05921033　系級：　電機所碩一　姓名：余政穎

1.請簡明扼要地闡述你如何抽取模型的輸入特徵（feature）

答：

在 hw1.sh 之中，我用的 features 抽取方式與助教相同，便是將每個月的 18 項 features 串起來，形成一個較大的矩陣，故每個月便有 471 筆訓練資料可以使用，12 個月共有 5652 筆訓練資料；Features 處理上，則是將連續九小時的 features 形成一筆訓練資料，並拉成直的，成為(指標數×9)維度的 features，最後將每一筆資料依序排列，使其成為一個矩陣，最後再使用 linear regression。(使用 gradient descent, eta=0.03 跑 10000 iterations)

在 hw1_best.sh 中，我抽取的 features 方法如下：

(1)先將 training data 讀入，除了直接使用每一個 feature 的值，亦將每一項的值平方與立方，並暫存入一個暫存矩陣之中，此時還沒有將每九個小時的 features 取出，只單純看某項指標的一次、二次、三次方，將可選擇的指標變為 3 倍。

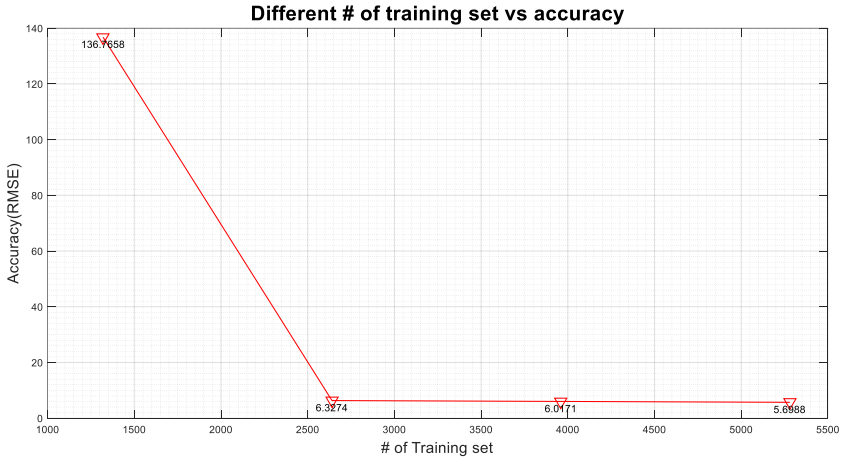
(2)接著，將上述指標與 PM2.5 的一次方比較正相關性(或者說是相似性，這邊用 cos similarity)，只選擇正相關性夠高的指標，將其連續九個小時的 features 取出，使一筆資料的維度為(所選指標的總數)×9，加上總共有 5652 筆資料，可形成一個矩陣。

(3)用上述矩陣去進行 linear regression。(直接以向量微分算出 w 最佳解)

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：

此題主要是藉由改變不同訓練資料量，去比較預測準確率的結果，此部分的實驗是分別將三個月、六個月、九個月及全部 training set 放入模型訓練，選擇 hw1_best.sh 中所採用的 features，正規化的參數λ設為 0，並上傳至 Kaggle，觀測上傳之後的 public set 之 root mean square error (RMSE)之結果：



可以觀察到隨著 training set 的數量減少，預測的結果會越來越不準，甚至當資料只剩下原本的 1/4 時，RMSE 升高至超過 100，linear regresstion 的預測已經完全不準確。

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：

本題比較分別將訓練資料的最高次方設定為一次方、二次方、三次方後，所得到的結果；第一種為將所有 features 丟下去 train，第二種則是利用第一題中抽取 features 的方法，

將挑選過的 features 丟下去 train，同時以 Kaggle 上的 Public data 之 RMSE(使用全部資料)與自己切一個 validation set 之 RMSE 為衡量標準，validation set 是取每個月的最後 31 筆資料集合而成，故放入模型訓練的 training set 資料量為 5280 筆，validation set 資料量為 372 筆，正規化參數皆設為 0，所得到的結果如下表：

	一次方	二次方	三次方
第一種(Kaggle public RMSE)	5.95828	5.99185	6.20186
第一種(validation set RMSE)	5.02747	5.04367	5.31840
第二種(Kaggle public RMSE)	5.92334	5.81294	5.69882
第二種(validation set RMSE)	5.20195	5.12687	5.32656

可以觀察到，再使用第一種設定，較簡單的 model 效果較好，可能是因為沒有特別挑選 features，所以有些高次方的指標反而預測的較不準確；然而，若有用計算相關性的方式，挑選適合的 features，因為較複雜的模型可挑選的比較多，則能得到較佳的預測結果。

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

答：

本題比較設定不同的正規化之參數λ，對於預測結果之影響；這題選擇 hw1_best.sh 中所採用的 features，使用資料的方式與第三題的設定相同：

	$\lambda = 0$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$
Kaggle public RMSE	5.69882	5.70668	5.71309	5.71844	5.72300	5.72695
validation set RMSE	5.32656	5.31591	5.31379	5.31237	5.31109	5.30986

可以觀察到，加上正規化的模型，其預測準確率有些影響，然而有趣的是，在 Kaggle 上與我自己的 validation set 所得的結果剛好相反，前者正規化越多準確率越差，後者正規化越多準確率越好，這可能是跟餵入的訓練資料之分布有關，Kaggle 上的 public set 的資料與 training set 較為相似，故做正規化反而看起來不準確，validation set 則相反。

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 $\sum_{n=1}^{N}(y^n - wx^n)^2$。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：

$$\text{The target is to minimize } E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(y^n - wx^n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(y^n - (x^n)^T w)^2 = \frac{1}{N}\left\| \begin{matrix} y^1 - (x^1)^T w \\ y^2 - (x^2)^T w \\ \dots \\ y^N - (x^N)^T w \end{matrix} \right\|^2 = \frac{1}{N}\|Xw - y\|^2 \quad (\text{N} \gg \#\text{ of features})$$

$$\nabla E_{in}(w) = \frac{1}{N}(2X^T Xw - 2X^T y) = 0 \Rightarrow w = (X^T X)^{-1} X^T y$$

$$In\ addtion, w_{reg} = (X^T X + \lambda \mathrm{I})^{-1} X^T y\ where\ \lambda \text{ is the regularization parameter}$$