學號：B02901093 系級： 電機四 姓名：吳岳

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

原本的做法是取18*9個小時的feature去得到第十個小時的PM2.5, 但在實作時發現成果不佳、運算也過於複雜。後來想到老師在解釋寶可夢的例子時用到的假設：我們一開始不知道他們的feature跟CP值有什麼直接的關係，但內行人都知道，有些feature (ex: 身高、體重)根本是雜訊，取了反而會增加誤差。

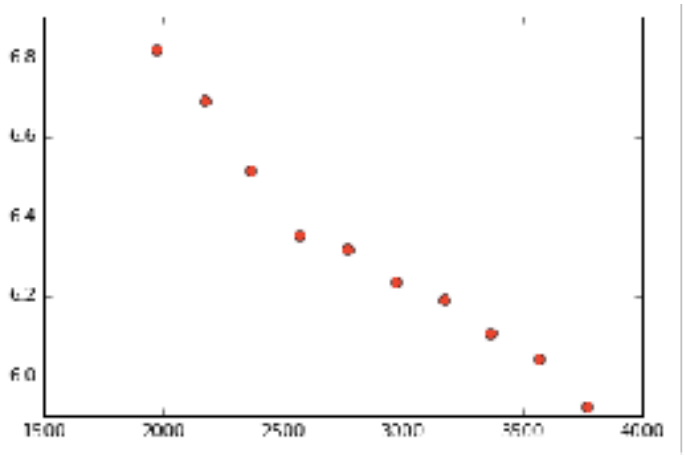
所以我決定只用最直接的feature — PM2.5去訓練，用每9筆的PM2.5資料去預測第十個小時的數值。

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：

這裡的準確率是依據Cross Validation去量測，我將2/3的資料作為training set, 剩下1/3的testing set 來衡量準確率。

在只考慮PM2.5的狀況下，資料總共有5652組, training set最多為5652*2/3 = 3768，我讓資料每次遞減200筆，取10筆資料來計算準確率。(3768, 3568…1968)



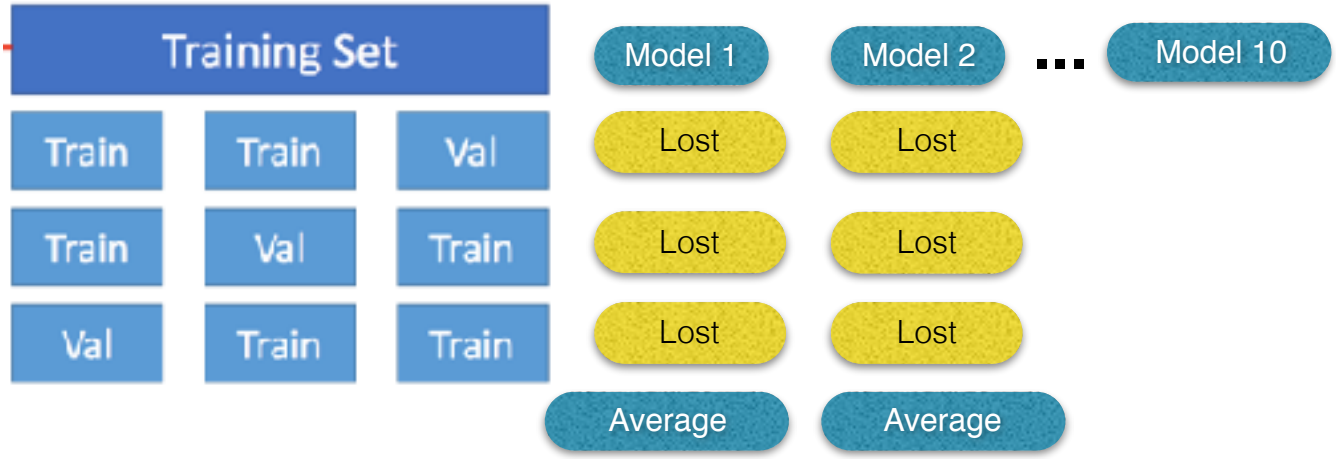
圖示的x軸為training 資料量, y 軸為RMSE, 當資料還維持在3768筆時，RMSE還可以維持simple baseline的水準(5.92)，而當資料剩1968筆時，RMSE已達6.819，可見不同資料量對準確率的影響。

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：

這邊比較兩種模型型態:

1. 一般: 任意給10種model, 用全部的資料來train, 取lost function最小的
2. Cross Validation: 資料分3組, 其中兩組當training set, 剩下一組當validation set, 排列一下發現有3種資料分類的狀況(如圖示)。任意給10種model, 取平均lost function最小的 model



最後可發現一般狀況比 Cross Validation 還要準確。由於一般狀況是將整個training set的資料都拿進去train, 判斷誤差時也是考慮既有訓練過的資料。而 Cross Validation 每次只考慮2/3的training set, 判斷誤差時是根據未見過的資料。故此 Cross Validation較不精準

然而，等真正將資料放進Testing Set時, 一般狀況在public data的的error rate很有可能會比Cross Validation大，因為一般狀況的精準性可能來自於Overfitting public data，而未考慮private的準確率。

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

答：

隨著regulation參數lambda的變化, 在testing set呈現出的error會有先增後減的趨勢，lambda的增加可以讓函式更smooth而避免overfitting，但太大的lambda又會讓函式過於簡單而讓error暴增。

當然，training set的error則無論如何都會隨lambda的變化增加，畢竟是在既有的解上直接給lambda變量。

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 xn，其標註(label)為一存量 yn，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n=1Nyn-wxn2 。若將所有訓練資料的特徵值以矩陣 X = [x1 x2 … xN] 表示，所有訓練資料的標註以向量 y = [y1 y2 … yN]T表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

答：

將任意矩陣T的轉置矩陣定義為T’。首先，根據迴歸線的矩陣形式: $X'Xw = X'Y$，這時在兩邊同時乘上X’X的反矩陣 ： $(X'X)^{-1}X'Xw = (X'X)^{-1}X'Y$

$(X'X)^{-1}X'X = I$ 消掉 => $\mathbf{w= (X'X)^{-1}X'Y}$