學號：B03902059 系級： 資工三 姓名：紀典佑

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

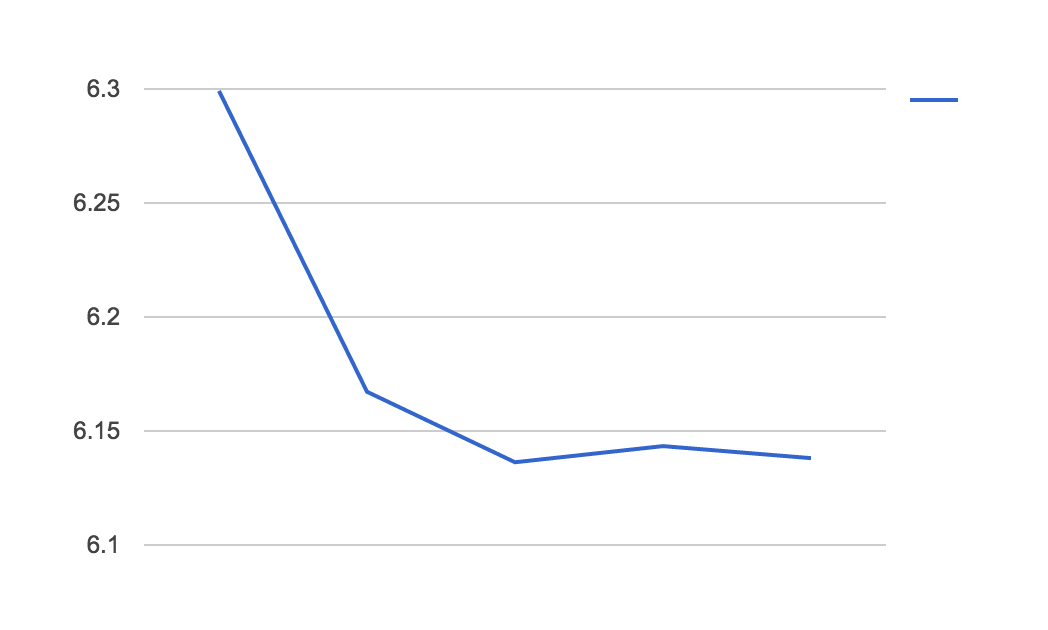
答：

一開始先抽取9\*18個完整的feature來train，之後用相關係數取出

所有feature中大於0.5的幾個來train，也做過各個指標與PM2.5的相關係數取出大於0.4的來train,而最好的結果(public)是拿七小時的PM2.5 train出來的。

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

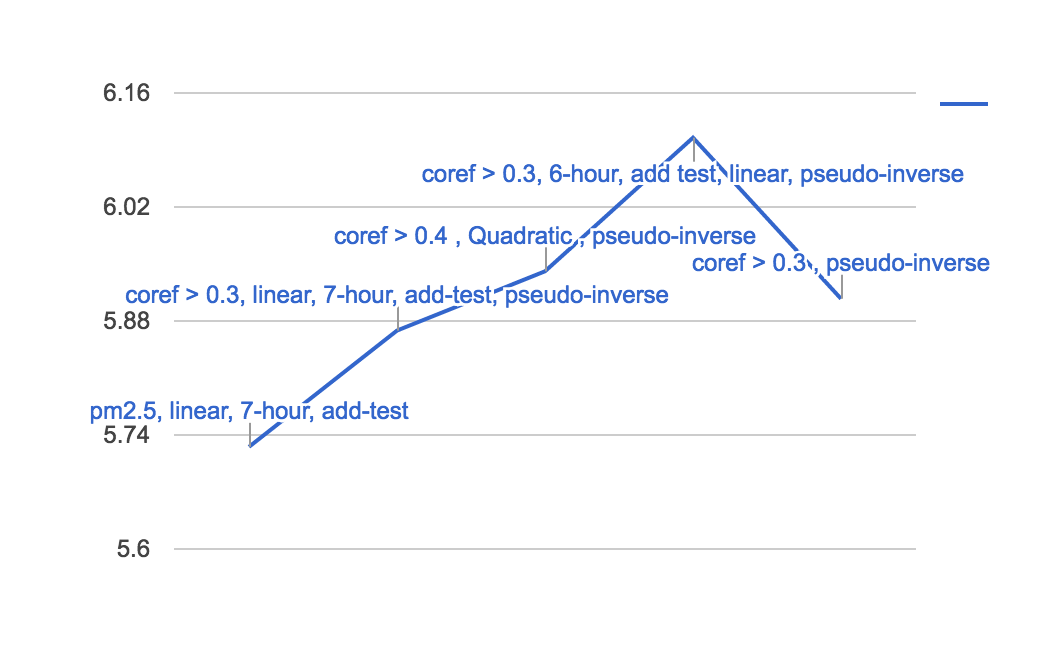
答：



各個折點分別為1000,2000,3000,4000及全部資料所測出來的validation error，使用的model為與PM2.5的相關係數大於0.4的幾個指標當做feature的結果，正常的情況下，資料量愈大，train出來的結果error會比較小，但是若是在多加入的data有noise的狀況時，可能會使得error上升，圖表中1000與2000比之間的差距可能為前者，而後折線的浮動原因可能為後者。

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：



資料複雜度由左而右分別為：

1. 只取PM2.5七個小時的feature，並且將test data切割七小時取出併入training data裡頭，做一次方的linear regression。
2. 取指標與PM2.5的相關係數大於0.3的七個小時feature，一樣加入七小時的test data，做一次方的linear regression。
3. 取指標與PM2.5的相關係數大於0.4的九個小時feature，做二次方的linear regression。
4. 取指標與PM2.5的相關係數大於0.3的六個小時feature，加入六小時的test data，做一次方的linear regression。
5. 取指標與PM2.5的相關係數大於0.3的九個小時feature，直接做一次方的linear regression。

以上除了第一個沒用最佳解外其他皆有使用，可以直接判斷該模型能到達Gradient descent的最低點，取相關係數高的原因是為了讓一些與結果較無關的資料不要加入training過程，但取相關係數高的feature做出來的結果並未比只取PM2.5的結果好，另外取相近的幾個小時的結果也有些浮動，我推測的原因是因為PM2.5在空氣中存在的時間有關，另外使用二次的結果比一次的差一點，可能發生了overfitting，因為在整體的training data error的結果，二次是比一次好的。

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

答：

正規化主要是為了避免training的結果overfitting, 而發生overfitting的主要原因是選取model的degree過大，使得結果接近於training data，若training data的noise過多，就會讓結果跟test相去甚遠。

這次的作業中，我實作了二次的linear regression，加入regularization並且調整lambda，lambda變大，在整體training data的error會稍微變大，但是將得到的結果放上Kaggle做測試，並未得到比一次linear regression小的error，我推測可能的原因是這次資料的feature與結果的關係，比較適合使用一次model來training。

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 。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：

=>

min -> =>