學號：R05921078  系級： 電機碩1  姓名：洪立達

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

參考一些空氣污染的研究報告去掉一些可能無關的因素，之後再嘗試搭配某些因素取次方向來作測試。最終取了AMB\_TEMP、CO、NO2、NOx、PM10、PM2.5、PM2.5\*\*2、SO2、WD\_HR、WIND\_DIREC為我的feature

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：

|  |  |  |  |
| --- | --- | --- | --- |
| Hours | Degree | Times | Loss |
| 4 | 2 | 5000 | 8735.158 |
| 4 | 2 | 7000 | 8738.07 |
| 4 | 2 | 9000 | 8738.067 |
| 4 | 3 | 5000 | 9096.855 |
| 4 | 3 | 7000 | 9162.037 |
| 4 | 3 | 9000 | 9220.128 |
| 5 | 2 | 5000 | 8645.364 |
| 5 | 2 | 7000 | 8665.507 |
| 5 | 2 | 9000 | 8675.121 |
| 5 | 3 | 5000 | 8917.751 |
| 5 | 3 | 7000 | 8953.3 |
| 5 | 3 | 9000 | 8981.086 |
| 6 | 2 | 5000 | 8742.454 |
| 6 | 2 | 7000 | 8792.784 |
| 6 | 2 | 9000 | 8833.6 |
| 6 | 3 | 5000 | 8875.926 |
| 6 | 3 | 7000 | 8943.304 |
| 6 | 3 | 9000 | 8994.508 |
| 7 | 2 | 5000 | 8909.323 |
| 7 | 2 | 7000 | 8965.619 |
| 7 | 2 | 9000 | 9008.55 |
| 7 | 3 | 5000 | 9270.323 |
| 7 | 3 | 7000 | 9421.393 |
| 7 | 3 | 9000 | 9507.614 |

Hours：一次取多少小時的資料區間做為feature

Degree：feature的複雜度

Times：訓練次數

Loss：用Test\_X前八小時來預測第九小時的pm2.5之誤差

(圖表為我在測試提升feature複雜度與訓練次數是否對預測準度有所幫助)

由圖表發現，訓練次數達到7000次後，誤差值大致上也趨於穩定，不過取的時間範圍越大，訓練次數增加時，誤差值的增加也變大。猜測的原因是使用adagrad後，訓練次數到達一定數量時，就會趨於穩定。另外我也測了一次極端的情況，調整次數為三萬次，不過依舊對準確率沒有進一步的幫助。

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：

(同上圖)

在測試調高複雜度後，出來的loss值確實比一次方的時候來得好，但是上傳kaggle的結果確是相反，推測是已經overfitting了。因此最後我只取了幾個感覺比較重要的feature加入二次式來測試，出來的結果是我目前的best case。

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

答：

在提高複雜度後有可能出現overfitting的情況，加入regularization可以讓高次方的曲線較為平滑，進而提升準確率。不過我自己的測試是加入regularization後，準確率還是不如原本一次方的model，所以最終沒有使用regularization。

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：

w =