學號：B03705027 系級： 資管三 姓名：鄭從德

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

我一共抽取了八個觀測值：分別是pm2.5、pm2.5的平方、SO2、O3、風速、風向、一小時平均風速(WS\_HR)、一小時平均風向(WD\_HR)。這些都是取前九期的資料，所以一共有72個features。我的選擇方法是每個側項都各別單獨與pm2.5跑一個model出來，error下降的我就會在我最後的model中選擇它。

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

這是我用Kaggle上的public set 得分與dataset大小關係做的圖。我僅挑model接近的幾筆資料進行做圖，因此資料點較少。由圖中可以看出，在我挑選training data多的時候，每次Error較近，變異數較小；後來利用random的方式切小training set，error的變異數因此增加，分布的更分散。總題而言還是小的training set 誤差比較小。我覺得可能是因為取較多資料的話，同個資料點會被取到很多次，可能因此造成overfit。

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響

若我們以feature數目判斷複雜度，則基本上feature越多error越低。

若我們以方程式複雜度討論，則由實作結果可以歸納：PM2.5取二次方有助於error降低；其他的features的二次式則沒有顯著較果。三次方時則會使test error飆升(overfit)，因此我在最後model中僅使用二次式的模型。

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

上圖為重建一組model並隨機的進行train test split後的做圖結果。 testing set最低誤差出現在 lambda = 6 的時候。再加入lambda的過程中，testing error有變小的趨勢(幅度可達0.3)，但過了一個臨界點後，testing error還是會上升。

結論：適當的挑選lambda值，真的可以降低testing error!!

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 。