學號：R05921064 系級： 電機碩一 姓名：羅嵩詠

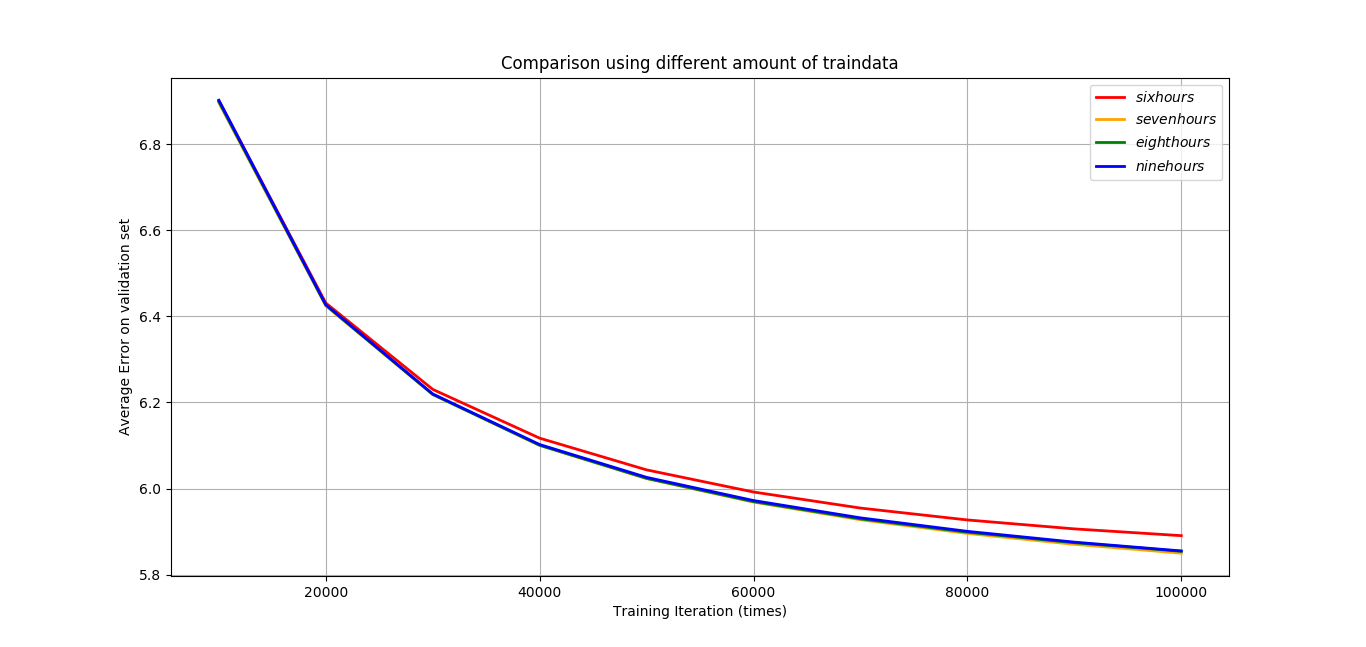
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

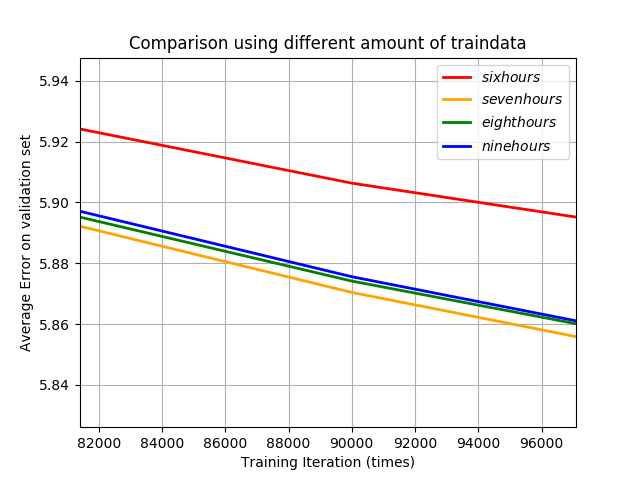
答：考慮train.csv提供的18種會影響PM2.5含量的因素(包括PM2.5在內)，並再根據簡單的推理(有用Kaggle驗證過)--”越久以前的data對於下個時刻的預測越不相干”，選取前七小時作為參考依據。故每筆輸入特徵長度為7x18=126。從每月第一天開始，相鄰的七個小時會串接成為一筆輸入特徵，最後一筆為第二十天的16點到22點。因此，每個月有479-7+1=473筆，乘上12個月，最後得到輸入特徵陣列大小為(126，5676)。

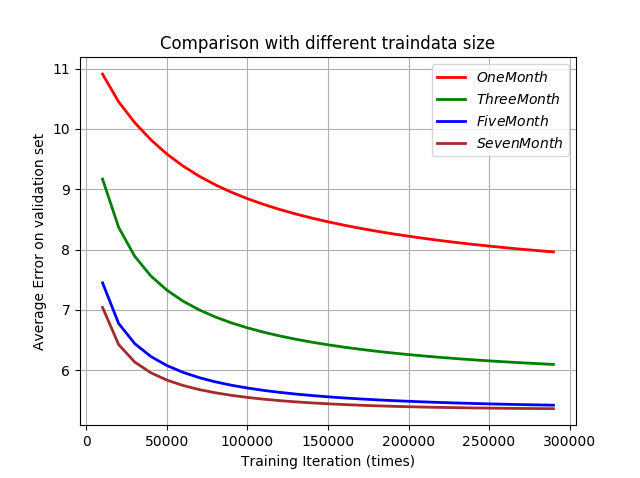
2.請作圖比較不同訓練資料量對於PM2.5預測準確率的影響

答：用的模型是一階的y=b+wx。

第一種比較方式，分別使用前六、前七、前八、前九小時，觀察對PM2.5濃度的預估會有什麼影響。

訓練10萬次後，由圖形可發現只用前六小時誤差明顯大於其他三者，只用前八小時與前九小時全用相當接近，而只用前七小時的RMSE最小。



第二種比較方式，只取1月、取1-3月、取1-5月、取1-7月的traindata。結論是用越多traindata在觀察範圍內誤差可修正到越小。誤差由取一個月到取七個月依序為:7.962108、6.096783、5.421128和5.362943。

從右圖可觀察到取的月份數越少，收斂速度

較慢，斜率從頭到尾變化都平緩，bias也比

較大。

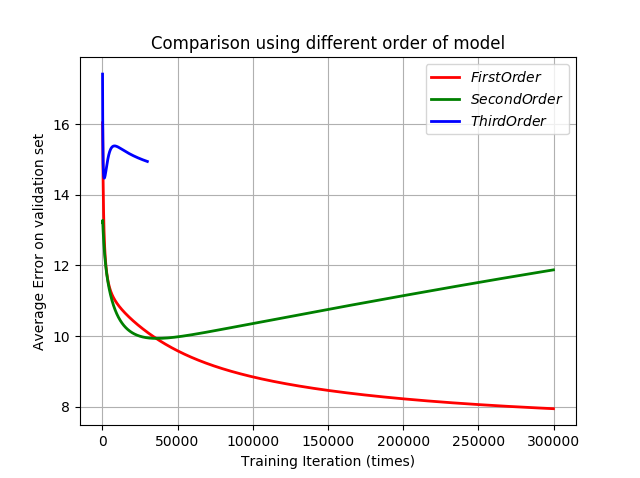
3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響

答：實作兩種複雜度：一維、二維和三維模型(尚未加入regularization term)

(為了降低多維模型計算量，取一月的data作為輸入特徵，九到十二月作為validation set)

在一維的情況下，訓練次數為30萬次，Minimum average error=7.945269

在二維的情況下，訓練次數為30萬次，Minimum average error=9.935060

在三維的情況下，訓練次數為3萬次，Minimum average error=14.474549

多維在預測精準度上，variance對其影響

較大，曲線呈現先降後升。而且學習率η

設定要小。三個模型中，只有一維模型在

訓練30萬次內會收斂。不過因為只使用

一個月的traindata，即使訓練30萬次，

誤差依舊很大。若data size增至8個月

，誤差可降至。

學習率η

一維 -> E-10 二維 -> E-13 三維 -> E-18

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

答：正規化在loss function加上一項模型參數的平方和，目的是希望讓模型平滑化。

這裡使用三維模型，訓練次數為訓練次數為3萬次，λ有0.001, 0, 1000等。未觀察到變化。

