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）

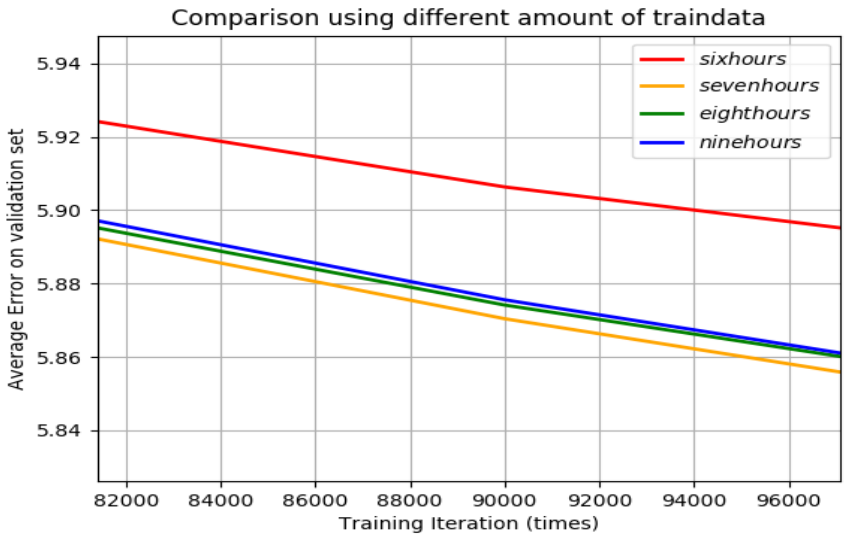
答：考慮 train.csv 提供的 18 種會影響 PM2.5 含量的因素(包括 PM2.5 在內)，並再根據簡單的推理(有用 Kaggle 驗證過)--”越久以前的 data 對於下個時刻的預測越不相干”，選取前七小時作為參考依據。故每筆輸入特徵長度為 7x18=126。從每月第一天開始，相鄰的七個小時會串接成為一筆輸入特徵，最後一筆為第二十天的 16 點到 22 點。因此，每個月有 479-7+1=473 筆，乘上 12 個月，最後得到輸入特徵陣列大小為(126，5676)。

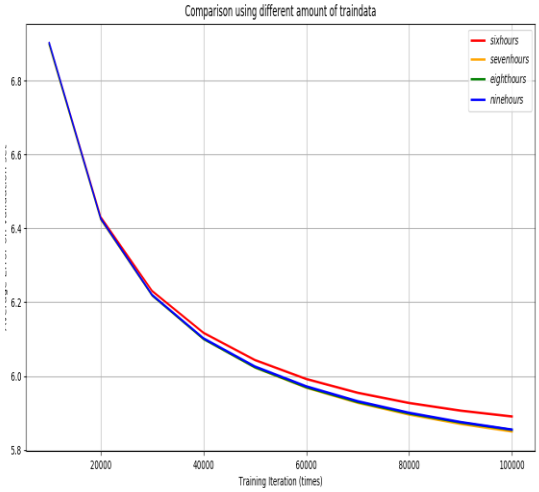
2.請作圖比較不同訓練資料量對於 PM2.5 預測準確率的影響

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。

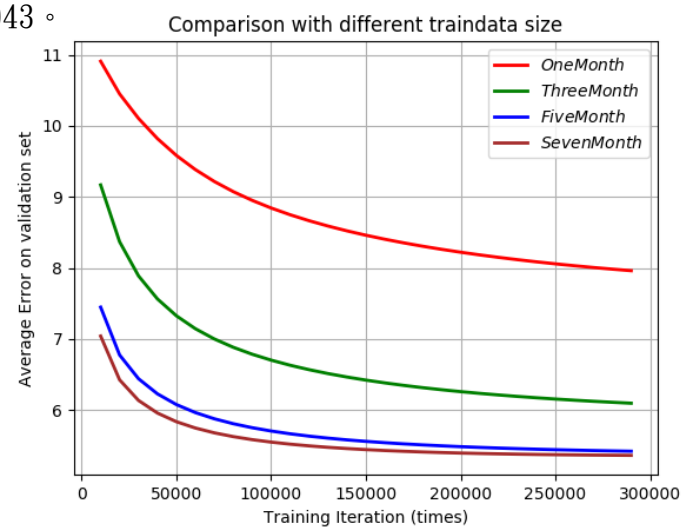
訓練 10 萬次後，由圖形可發現只用前六小時誤差明顯大於其他三者，只用前八小時與前九小時全用相當接近，而只用前七小時的 RMSE 最小。





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

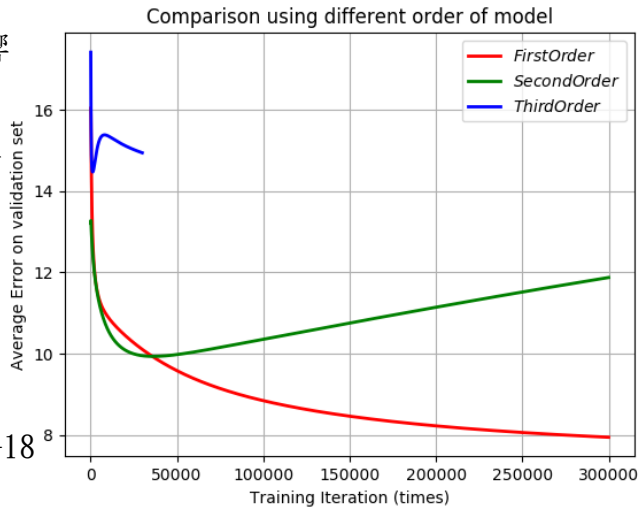
在二維的情況下，訓練次數為 30 萬次，Minimum average error=9.935060

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。