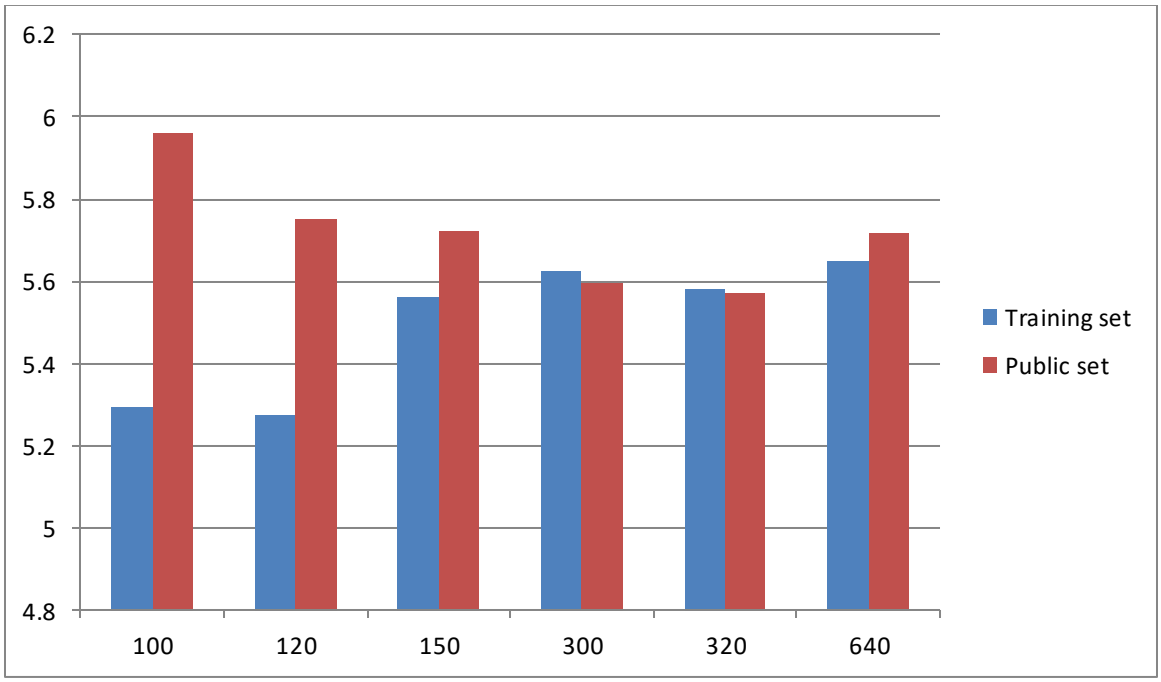
學號：R05921077 系級： 電機碩一 姓名：陳立杰

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：
我在 Kaggle 上最後選兩種抽法，第一種是抽預測前八個小時的 PM2.5 共 8 個 feature，第二種是抽預測前八個小時的 PM2.5+PM10 共 16 個 feature，交上來的檔案是 8 個 feature 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：



由圖表可以看到當資料量越少的時候 training set 的 error 就會越小，但在 public set 的部分卻是資料越多時 error 會比較偏小，但不完全成正相關，像是 public set 在 320 筆資料的時候 error 就比 640 筆資料時小很多。

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：
我用前八個小時的 PM2.5 來預測在 training set 的時候 error 為 5.58224702012，在 public set 為 5.57012。另外用前八個小時的 PM2.5+前一個小時的所有資料在 training set 跟 public set 預測的 error 為 5.250937280675 和 5.89358，雖然說模型複雜度提升時 error 可能會變少，但是也有可能 overfitting 了，所以在 testing set 測的時候 error 不見得會比較少。但如果用太過於簡單的模型來預測也會增加錯誤率，像我只用前兩個小時的 PM2.5 來預測時 training set 的 error 為 6.58187030796。

4. 請討論正規化(regularization)對於 PM2.5 預測準確率的影響
答：
我用前八個小時的 PM2.5 來做預測原本的 training set 跟 public set 的 error 為 5.58224702012 和 5.57012 加上了 regularization 後，當 lambda=1 時為 5.59239659748 和 5.57048，當 lambda=10 時為 5.87929213229 和 5.80878，從這些數據分析出有做正規化的時候在 public set 的表現相較 training set 會比沒有做正規化的時候比較好，lambda 較大一點的話就會越明顯，但是加入正規化之後 training set 的表現就會變差，所以整體來看沒加正規化時的 public set 表現還是比較好。

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 x^n，其標註(label)為一存量 y^n，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$\sum_{n=1}^{N}(y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \dots x^N]$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \dots y^N]^T$ 表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。
答：
w = 2*Xy