機器學習 Homework #1 Report

學號 - **B03902015**

系級 - 資工三

姓名 - 簡瑋德

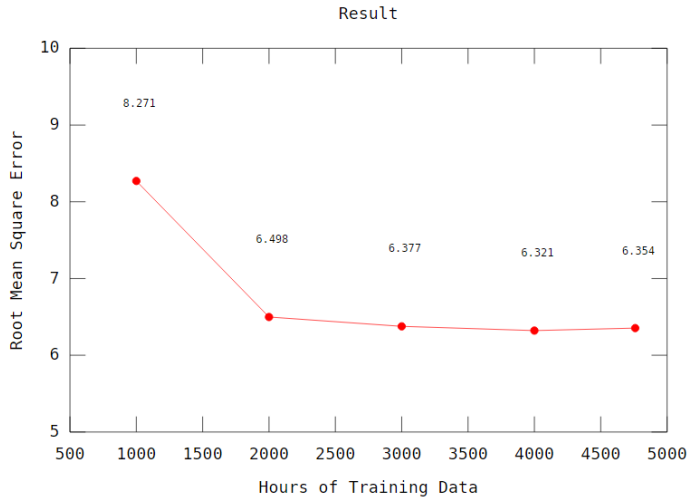
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

- 考慮前9天的資料，每天有18項觀測記錄，一共162維
- 因有些「RAINFALL」的觀測數值為「NR」，以「0.01」取代之
- 對資料做Feature Scaling，將各維數值的範圍縮放到$[0, 1]$

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

- Training Dataset中一共有5760小時的觀測紀錄，將最後的1000小時預留，作為Validation Dataset
- 以Root Mean Square Error作為準確率的指標

Traing Data (Hours)	E_training	E_validation
1000	5.764	8.271
2000	6.011	6.498
3000	6.251	6.377
4000	6.052	6.321
4760	5.944	6.354



- 由圖可見，訓練資料增加，有助於預測準確率的提升

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響

- Training Dataset中一共有5760小時的觀測紀錄，將最後的1000小時預留，作為Validation Dataset，其餘4760小時拿來訓練
- 以Root Mean Square Error作為準確率的指標
- 放棄時距較遠的觀測資料，試著降低模型的複雜度

Feature Dimensions	E_training	E_validation
162	5.767	6.194
18	6.418	6.742

- 由表可見，若使用複雜度太低的模型，可能會面臨「Underfit」的問題，模型沒辦法完整地擬合訓練資料，進而影響準確率

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

- 從上述實驗的數據中，可以發現，即使將前9個小時的數據全部抽取出來當作Feature，Traning和Validation的Error都相差不大，因此沒有「Overfit」的問題，正規化並無法有效提升準確率
- 當然，如果非得使用太過複雜的模型，使Training Error降到很低但Validation的結果卻很差勁，這時候就可以考慮正規化，避免Overfit

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)為一純量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$\sum_{n=1}^{N}\left(y^{\mathrm{n}}-\mathrm{w}\cdot\mathrm{x}^{\mathrm{n}}\right)^2$。若將所有訓練資料的特徵值以矩陣$X=[\mathrm{x}^1\mathrm{x}^2\ldots\mathrm{x}^{\mathrm{N}}]$表示，所有訓練資料的標註以向量$\mathrm{y}=[y^1\ y^2\ldots y^{\mathrm{N}}]^{\mathrm{T}}$表示，請以 X 和 y 表示可以最小化損失函數的向量 w

- $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{y}$