Machine Learning Project 1

PM2.5 Predictor

Team Member

- b03902089 資工三 林良翰

Best Result

- Iteration: 10000
- Learning Rate: 1.25×10^{-8}
- Using Time: 7 hours
- Using Feature: O_3, $PM2.5$, SO_2
- E_{in}: 6.0755
- E_{out}: 5.7922

Questions

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
    - 將資料存成一個18×5760的二維陣列，每一列(row)代表各種不同的輸入特徵(feature)，每一行(column)代表每一個小時的資料，因此這個資料在讀入的時候每讀十八行(line)都要做一次處理，讓資料的時間是全部連在一起的，而不是被分割成24小時為單位。
    - 在讀入降雨量(RAINFALL)的時候，有些值是NR(NO RAIN)，因此要把這個值轉換成0.0浮點數。
    - 在讀入PM2.5的時候，有些值是−1，要把它轉換成0。
    - 計算每個特徵與PM2.5之間的相關係數，並篩選出對PM2.5有比較高影響力的特徵。
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
    - 總資料量：5760筆
    - 使用特徵：PM2.5
    - Validation：最後576筆資料
    - Learning Rate：10^{-8}
    - Iteration：5000
    -

Data Size	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
E_{in}	6.11	6.14	6.07	6.25	6.52	6.05	6.38	6.38	6.25
E_{valid}	7.14	7.13	7.14	7.13	7.12	7.18	7.20	7.27	7.58

- 由上面的表格可以知道資料量太少的話準確率會降低。

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響

- 當只使用PM2.5作為訓練特徵時，其E_{out}就可以達到5.91左右。
- 當使用O_3、PM2.5、SO_2作為訓練特徵時，E_{out}降到5.78左右（我目前的最佳成果）。
- 當使用NO_2、O_3、PM10、PM2.5、SO_2作為訓練特徵時，E_{out}又回提升到5.86。
- 最後當使用全部的特徵拿去訓練時，E_{out}會衝到6.0以上。

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

- 總資料量：5760筆
- 使用特徵：PM2.5
- Validation：最後576筆資料
- Learning Rate：10^{-8}
- Iteration：5000
-

λ	0.01	1.0	100.0
E_{in}	6.1059	6.1069	6.1069
E_{valid}	7.1432	7.1432	7.1429

- 正規化影響不大

5.
- 損失函數L：
$$L = \sum_{i=1}^{N} (y_i - x_i w)^2 \ = (y - Xw)^2$$
- 為了求L的最小值，經過微分(對w)後可以得到：
$$\frac{\partial L}{\partial w} = 2X^T(y - Xw) = 0$$
$$2X^T Xw = 2X^T y$$
- 如果$X^T X$有反矩陣，最後可得：
$$w = \left(X^T X\right)^{-1} X^T y$$