Homework 1 Report - PM2.5 Prediction

B05902013 資工二 吳宗翰

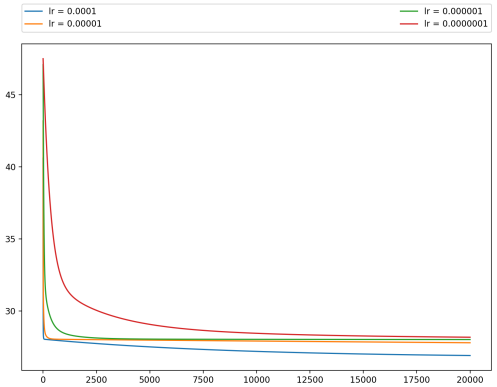
1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

Table 1: 下表是用全部參數與 pm2.5 訓練的結果

feature	E_{in}	E_{out} (public)	E_{out} (private)
9+bias	27.99	13.68192	13.28569
162+bias	25.34	11.55014	10.99377

本次比較了 10 個參數和 163 個參數 train 出來的結果，使用 Gradient Descent，$\mathrm{epoch} = 20000, \mathrm{lr} = 0.0001$。乍看之下從表中來看是 163 個參數的比較優就會覺得是因為 10 個參數可能還不夠多，因此擁有 163 個參數的模型跑出來的結果較佳。然而實際上的狀況卻是因為本次的資料 pm2.5 這一項有蠻多不正確的資料，因此在沒有 preprocessing 的狀況下，單獨使用 pm2.5 的表現比較糟，而如果有 163 個參數則可以用其他參數補救。(在這題中我並沒有去比較 preprocessing 後誰比較好)

2. (2%) 請分別使用至少四種不同數值的 learning rate 進行 training（其他參數需一致），作圖並且討論其收斂過程。



上圖是第一小題 9 小時 pm25+bias 項做 Gradient Descent 的結果，橫軸是迭代次數 (共 20000 次)，縱軸是 E_{in}。另外本次選用的 learning rate 是 10^{-4} 到 10^{-7} 四種。

從收斂過程上來看，不意外的比較大的 learning rate 可以學得比小的還要快，正如圖中 10^{-4} 這組遞減的最為快速；然而另一方面 learning rate 也不能太大，像是我有實驗 10^{-3} 的 learning rate，不過因為 RMSE 大到完全無法收斂所以就沒放上來了。這題告訴我們在做 Gradient Descent 的時候，learning rate 也是要好好去挑的。

3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

Table 2: 下表是本次實驗 4 種不同 λ 所得出的結果

λ	E_{in}	E_{out} (public)	E_{out} (private)
10	26.91	12.87988	12.36921
10^2	26.89	12.87684	12.36296
10^3	26.79	12.89749	12.34367
10^4	228.42	181.72241	175.64521

本次使用 L2 regularization 的相關設定是：$\text{epoch} = 20000$ 加上 Gradient Descent，$\text{lr} = 10^{-4}$，選用的參數是前 9 小時的 pm2.5 加上 bias 項。

從實驗結果中可以發現，在本次訓練中 λ 不大的時候並沒有對訓練結果產生具體的影響。不過可以明顯地看到當我們不小心把 λ 設太大的時候就會因為 regularize 那一項 dominate 而使得訓練壞掉，我想這是我們日後在做 regularization 的時候值得警惕的地方。

4. (1%) 請這次作業你的 best_hw1.sh 是如何實作的？(e.g. 有無對 Data 做任何 Preprocessing？Features 的選用有無任何考量？訓練相關參數的選用有無任何依據？)

這次的 best_hw1.sh 實作上大致可以分成三個比較重要的點來看：(1)Feature 的選取 (2) 資料處理 (3) 訓練模型，在以下我將簡單的敘述一下做法：

(1) Feature 選取

在這次作業中，我選取了前九個小時的 PM2.5 以及 PM10 作為 feature。選取這些參數的原因是因為發現他們彼此之間的相關係數蠻大的，因此就優先考慮這些參數；至於取前 9 個小時是實驗出來的，畢竟我其實也是很怕參數太多可能會 Overfitting。

(2) 資料處理

我覺得這次作業中能讓我 best_hw1.sh 能過 strong baseline 的原因就是有做資料處理。首先在看完 Training data 後，我發現大於 120 以及小於 2 的資料 (PM2.5, PM10) 似乎是不太合理，因此在我有先把他們從 Training data 中丟掉；至於 Testing data 我則是對他做插值，也就是用旁邊幾個點的平均來取代我覺得壞掉的點的資料。另外為了訓練隔夜的狀況，我有把原本的 240 天 $\times$ 24 小時變成 12 個月 $\times$ 連續 480 小時，我發現這個訓練結果也有變好。

(3) 訓練模型

訓練模型其實也是 Linear Regression 下去跑而已，並沒有什麼特別的，另外我也沒有加上 Regularize 的項。