學號:T10902303 系級：計資一 姓名：梁高浚

請實作以下兩種不同 feature 的模型，回答第 1 ~ 2 題：

1. 抽全部 9 小時內的污染源 feature 當作一次項(加 bias)
2. 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註：

a. NR 請皆設為 0，其他的非數值(特殊字元)可以自己判斷

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等)都是可以用的

c. 第 1 ~ 2 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好，kaggle 死線之後便可以無限上傳。

1. (1%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

(1)如果是使用全部 9 小時內的污染源 feature 作為一次項，RMSE 是 8.74189(public), 9.85006(private)

(2)如果是使用全部 9 小時內的 pm2.5 的一次項作為 feature, RMSE 是 8.03427(public), 9.11282(private)

這裡可以看出來第二種方法的誤差值比較小，原因是因為第一種方法可能出現了 overfit 的問題，使用過多的 feature 會令到模型過分貼合訓練集，當上傳到 kaggle 的時候碰見未知的 data 就沒法很好的 generalize 預測的結果。

另外，第二種方法其實算是時間序列的模型，其使用了前九小時的 pm2.5 值預測後面時間的 pm2.5 值，它的準確率比較高，可能說明了 pm2.5 這個變量本來就是比較 time-dependent, 就是說它會隨著時間的變動而變化，合理的解釋可能是上下班通勤時間車輛產生的廢氣、其他人為活動令到 pm2.5 上升。

2. (1%)解釋什麼樣的 data preprocessing 可以 improve 你的 training/testing accuracy，e.g., 你怎麼挑掉你覺得不適合的 data points。請提供數據(RMSE)以佐證你的想法。

如果是使用第一個模型，即使用 X1,X2,...,Xn 等 feature 預測 Y, 第一個可以做的 data preprocessing 是檢查 Xi <-> Y 的線性關係。因為在多元線性回歸裡面，其中一個重要的假設是 predictor variables 與 response variable 存在線性關係。

第二個重要的假設是檢查 Xi <-> Xj, i =/= j，因為在多元線性回歸裡面，另外一個重要的假設就是 predictor variables 之間是獨立不相關的，這樣的話(X'X)^(-1)才有解。