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。