ML hw1 Report

學號：B05705001 系級： 資管四 姓名：黃意芹

請實做以下兩種不同feature的模型，回答第 (1) ~ (2) 題：

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

(1) 抽全部9小時內的污染源feature當作一次項(加bias)

RMSE

(2) 抽全部9小時內pm2.5的一次項當作feature(加bias)

RMSE	全部9小時內的污染源當作feature	全部9小時內pm2.5當作feature
private	5.65158	5.72749
public	6.02616	5.99585
average	5.83887	5.86167

由得到的RMSE可知，抽全部9小時內的污染源feature當作一次項的表現會比抽全部9小時內pm2.5的一次項當作feature更好，不過差距不是很大。原因可能是因為PM2.5跟其他污染源有關聯，所以增加更多feature對於預測PM2.5會有所幫助。不過因為PM2.5本身就算是最主要的特徵，且其他資料可能有摻雜許多雜亂、異常的值，所以增加其他特徵的進步幅度沒有非常大。

2. 解釋什麼樣的data preprocessing 可以improve你的training/testing accuracy，ex. 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。

RMSE	全部9小時內的污染源當作feature且有清除掉異常pm2.5	全部9小時內的污染源當作feature沒有清除掉異常pm2.5
private	5.69725	5.65158
public	5.54239	6.02616
average	5.61982	5.83887

除了 NR皆設為0，還有把在字尾的其他符號去除只留數字，我還有做資料清理方法為：先把pm2.5的極端值挑出（pm2.5 < 2 or pm2.5 >= 100），之後算出pm2.5的平均和標準差，將平均+−3.6*標準差以外的training data丟掉。在predict時，若testing data有pm2.5的值在平均+−3.8*標準差以外，則取代成平均值進行預測。會

這麼做的原因在於，pm2.5本身為最重要的特徵，然而資料有一些異常值，所以先清除掉再訓練出的model會比較準確。而testing data因為沒辦法丟掉資料，所以就把異常值取代為平均值進行預測。

從上表RMSE可觀察到，有去除異常pm2.5的model表現明顯較好。這是因為訓練與預測時不會被異常值誤導，所以結果也會比較準確。

3. Refer to math problem