學號: B05602052 系級: 電機四 姓名:舒泓諭

請實做以下兩種不同 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。
- e. 根據助教時間的公式表示,(1)代表 p = 9x18+1而(2)代表 p = 9*1+1

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

	9x18+1	9x1+1
public	10.46342	5.96943
private	5.95145	5.89724

不論在 public score 還是 private score 中,9x18+1 的模型都來得比 9x1+1,要來的更高。從此可以發現 9x18+1 的模型,因為考慮了太複雜的 model 所以有 overfitting 的情形。

2. (1%)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。

我認為可以藉由挑掉極端值,以降低 RMSE(根據 kaggle public+private 分數)。以下都是使用 p = 9x1+1 的 model。

	沒有挑掉極端值的	有挑掉極端值的
public	7.26793	5.96943
private	6.83433	5.89724

不論在 public score 還是 private score 中,如果有挑掉極端值的模型都會比沒有挑掉極端值的 RMSE 來的更低。

此外,我也有對如何**挑掉極端值的範圍**進行實驗。在統計學中的定義,極端值上界為 $(Q3+1.5\times IQR,\infty)$,極端值下界為 $(-\infty,Q1-1.5\times IQR)$ 。在本次的 data 中,極端值上界為 $(72.6,\infty)$,極端值下界為 $(-\infty,-40)$ 。然而因為這次 PM2.5 正常數值最小為 0,所以下界定為 $(-\infty,-0)$ 。

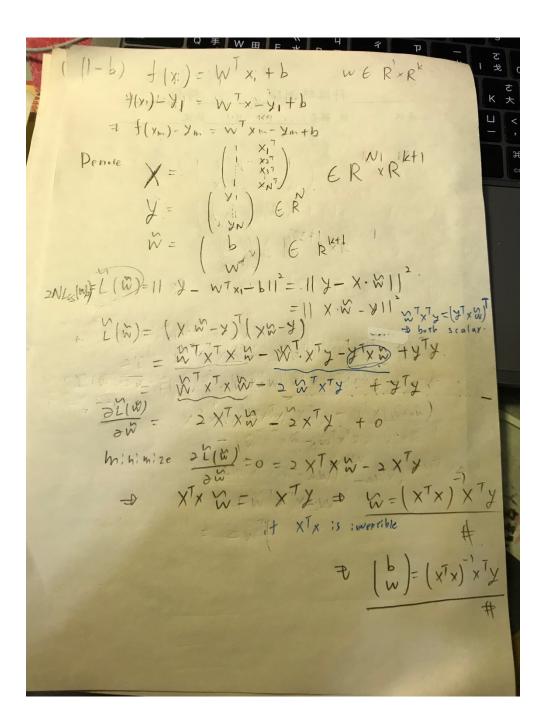
	range(2,100)以外挑掉	range(0,72.6)以外挑掉
public	5.96943	5.96926
private	5.89724	5.85901

從上可以發現雖然在 public score 這個根據統計學定義的極端值沒有什麼大幅的改變,但在 private score 卻上升了 0.04,可見藉由統計定義選出來的極端值,當所測資料量增加時效果才回比較明顯。

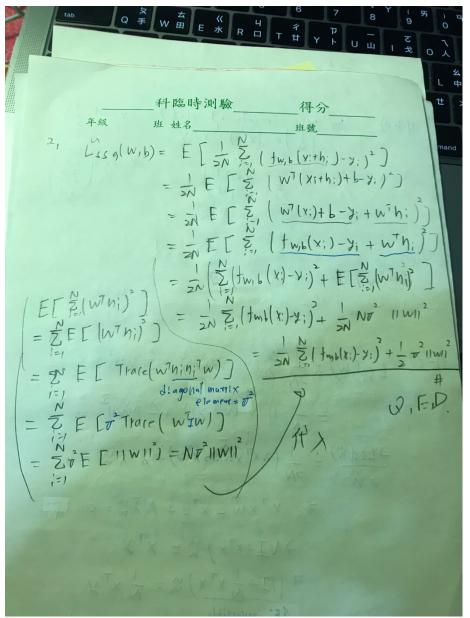
3.(3%) Refer to math problem

https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view

1. If y = y we result of y = (y = y) y



2.



3.collaborator 黃意芹

