Homework 2 Report - Income Prediction

學號：b05902002 系級：資工二 姓名：李栢淵

(1%) 請比較你實作的generative model、logistic regression的準確率，何者較佳？

gen.csv 18 minutes ago by Bai-Yuan Lee add submission details	0.84215	0.84508	☐
loge.csv 22 minutes ago by Bai-Yuan Lee add submission details	0.84731	0.85405	☐

個人實作的logistic regression的準確率比較好。

(1%) 請說明你實作的best model，其訓練方式和準確率為何？

我使用 sklearn 的 logistic regression，參數基本上影響不太大，主要是加了很多 feature，我將非 one-hot encoding的data ('age', 'fnlwgt', 'capital_gain', 'capital_loss', 'hours_per_week')，全部拿去疊，從二次方疊到五十次方再加上 sin、cos、tan 和 arctan，最後再一起 normalization，全部丟下去一起 train，就結束了。

public 的準確率為0.87678，private 的準確率為 0.87077。

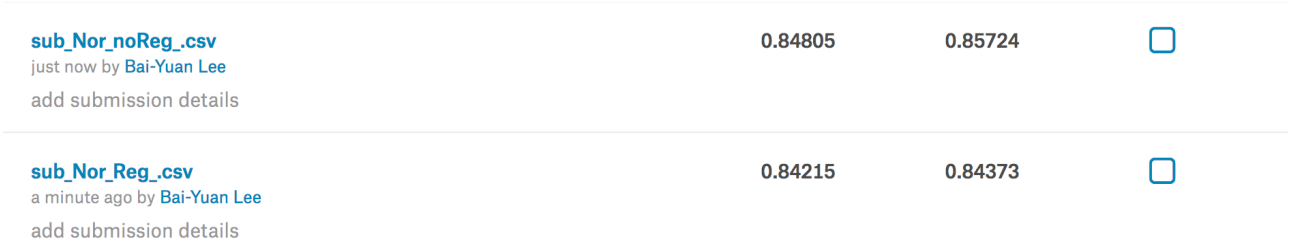
best.csv just now by Bai-Yuan Lee add submission details	0.87077	0.87678	☐

(1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關normalization請參考：https://goo.gl/XBM3aE)

sub_noNor_Reg_.csv just now by Bai-Yuan Lee add submission details	0.73958	0.74090	☐
sub_Nor_Reg_.csv a minute ago by Bai-Yuan Lee add submission details	0.84215	0.84373	☐

我初始 w 為 0 向量，有做 regularize，我發現有做 normalization 比沒做 normalization 好很多。可能是因為 trainX裡有五筆不是 one-hot encoding 的 data，尤其有些數據值又很大，如果沒做 normalize 就下去運算很容易overflow，也會讓那些非 one-hot encoding 的 data 的影響力比其他資料大太多，所以在這種情形下，我認為做 normalization 是很好的。

(1%) 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。(有關regularization請參考：https://goo.gl/SSWGhf P.35)



我初始 w 為 0 向量，有做 normalize，我發現有做 regularization 比沒做 regularization 差了一點，我認為是這幾筆資料問題，因為我做了很複雜的feature transform之後 private Score 還越來越高，代表說複雜的模型對這個是好的，而 regularization 有稍微簡化模型，所以覺得差了點很正常。

(1%) 請討論你認為哪個attribute對結果影響最大？

我覺得fnlwgt的影響最大，我一開始沒做normalization，覺得那數據又大又好像沒什麼相關性，所以把它拔掉再train，效果就變得很好。