學號：B05901025 系級：電機三 姓名：王鈺能

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

Accuracy	Public	Private	Total
generative model	0.84631	0.84117	0.84374
logistic regression	0.84778	0.84756	0.84767

Logistic regression 準確率較佳，我認為是因為 generative 用了一些不完全適用於此 model 的假設，像是 Output 是 Guassian(其實是 Bernoulli)、Covariance matrix 共用等等，導致無法較好的 fit 這些 data。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

用 keras 建一 3 層 NN，結構如下：

	unit	activation	regularizer
Layer1	100	relu	l1=0.001, l2=0.001
Layer2	80	relu	l1=0.001, l2=0.001
Layer3	1	sigmoid	None

取 10%training data 為 validation set，剩下以 batch size=32，train150 個 epoch

準確率為：

Public	Private	Total
0.8597	0.85702	0.85836

3. 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

我發現在沒有 normalization 的情況下 logistic regression trian 不太起來，lr=1、batch size=32，trian 50 epochs，他從頭到尾都卡在 accuracy= 0.2411 動彈不得；lr=100、batch size=32，trian 50 epoch，他也卡在 accuracy= 0.7589。normalization 後 training accuracy 就可以到 0.86 左右了，原因為 data range 差很多，像「fnlwgt」數量級為十萬左右，但其他像「education」等等採用 one-hot encoding，數量級就是 1，因此共用同樣的 learning rate 顯然是很難 update 的，必須 normalize 到差不多的 range 才比較好 update。

但 generative model 就比較不受 normalization 的影響，這是個合理的結果，因為它的原理是看每個 feature 在該維相對 mean 和 covariance 的分布關聯，和該 feature 數值大小無關，故不用 normalize 即可 train 起來。

4. 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。

r	Public	Private	Total
0	0.85331	0.85382	0.853565
0.0001	0.85343	0.85358	0.853505
0.001	0.85368	0.8537	0.85369
0.01	0.84449	0.84852	0.846505
0.1	0.76707	0.76047	0.76377

實作 l1 regularization，和作業一不同，這次 regularization 可以使準確率上升，在 r=0.001 時準確率最高，但更高又會 under-fit 了，我認為原因是這次模型不是 linear 的，和作業一為 convex 函數相比，此次的確可能發生 over-fit，所以適度的 regularize 可以讓 model 變好，準確率上升。

5. 請討論你認為哪個 attribute 對結果影響最大？

「fnlwgt」對結果影響最大，若在沒有 normalize 的情況下，由於它和其他 feature 數量級差太多，會導致 logistic regression 完全 train 不起來，但若把此 feature 拿掉，即使不 normalize 也可以過 simple baseline，可見他對結果的影響是相當顯著的。