Homework 2 Report - Income Prediction

學號：r05922080　系級：資工碩二　姓名：王鵬傑

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

Logistic regression 較佳:

Learning Rate: 0.01

EPOCH: 10000

	Public	Private
Logistic	0.85724	0.84829
Generative	0.80921	0.80764

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

Data Preprocessing:

(1) 將 train.csv, test.csv 的 workloss, education, marital_status, occupation, relationship, race, sex, native_country 做 categorical 調整。
(2) 將 age 個別調整，分成 0~30 歲、30~40 歲、40~50 歲、50~60 歲、60 歲以上的分類。
(3) 將 fnlwgt, capital_gain, capital_loss, hours_per_week 的值平方。
(4) 將 capital_gain, hours_per_week, education_num 做三次方。
(5) 所有參數做 Normalize。

Model:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 1024)              143360
_________________________________________________________________
activation_1 (Activation)    (None, 1024)              0
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800
_________________________________________________________________
batch_normalization_1 (Batch (None, 512)               2048
_________________________________________________________________
activation_2 (Activation)    (None, 512)               0
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 513
=================================================================
Total params: 670,721
Trainable params: 669,697
Non-trainable params: 1,024
```

3. (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)

	Public	Private
With Normalization	0.86031	0.85591
Without Normalization	0.80233	0.79253

因為我的 attribute 含有二次方及三次方項，因此若未做 normalization 每個參數的差異性會極大，導致訓練效果不佳，因此差距如此大。

但若拿掉平方項以及三次方項，準確率會更低，如下表：

	Public	Private
With Normalization	0.85651	0.85456
Without Normalization	0.77039	0.76796

4. (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)

Regularization term	Public	Private
$\lambda = 0.0$	0.85442	0.84743
$\lambda = 0.1$	0.85417	0.84768
$\lambda = 0.01$	0.85408	0.84741
$\lambda = 0.001$	0.85410	0.84743

在有 Regularization term 時雖然 Public score 略輸，但在 Private score 卻有較高分，且λ越小影響越小。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

Attribute	Public	Private
All attribute	0.86031	0.85591
Delete age	0.85405	0.84608
Delete workclass	0.85773	0.85050
Delete fnlwgt	0.85737	0.85394
Delete education	0.85589	0.85198
Delete education_num	0.85995	0.85763
Delete marital_status	0.85823	0.85480
Delete occupation	0.85049	0.84743
Delete relationship	0.85737	0.85382
Delete race	0.86044	0.85702
Delete sex	0.85921	0.85554
Delete capital_gain	0.83968	0.84117
Delete capital_loss	0.85700	0.85222
Delete hours_per_week	0.85601	0.85431
Delete native_country	0.85970	0.85652

從實驗的結果知道，race 是最不重要的，因此刪除他反而準確率變高，而 capital_gain 是最重要的，因為刪除他之後得到的準確率是最低的。