Homework 2 Report - Income Prediction

學號：b04901067 系級：電機三 姓名：陳博彥

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

Model	Public Score	Private Score
Logistic	0.85505	0.85503
Generative	0.76280	0.76474

Logistic的結果明顯較好，高了九個百分點左右（兩者皆有normalize，無regularization）。

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

Best的實作方式為：

1.所有attribute要normalize。

2.沒有regularization。

3.使用logistic model。

4.將continuous的attribute加上二次項及三次項。

5.batch size = 25, iteration = 3000, learning rate = 1

6.使用 np.clip(res,0.00000000000001,0.99999999999999) 來限制sigmoid的數值。

結果：Public Score: 0.85505 Private Score: 0.85503 出奇的接近！

另外，我也嘗試使用deep learning的方式實作，效果也不錯（但最後best不是選這個）。

karas參數如下：

```
model=Sequential()
model.add(Dense(500,input_dim=123,kernel_initializer='normal',activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(500,activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(500,activation='relu'))
```

```
model.add(Dropout(0.2))
model.add(Dense(1,activation='sigmoid'))
sgd=SGD(lr=0.01,decay=1e-6,momentum=0.9,nesterov=True)
model.compile(loss='binary_crossentropy',optimizer=sgd,metrics=['accuracy'])
model.fit(x_train,y_train,epochs=30,batch_size=150)
```

結果：Public Score: 0.85402　Private Score: 0.85518

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。

Model	Private Score	Public Score
logistic有normalize	0.85505	0.85503
logistic無normalize	0.85138	0.85264
generative有normalize	0.80862	0.79975
generative無normalize	0.79176	0.78774

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。

從數據得知，regulation沒有比較好，少數情況會爛掉。

lambda	normalize	無normalize
0	0.85505	0.780538
10^-4	0.85501	0.77934
10^-3	0.85489	0.763771
10^-2	0.85491	0.236225
10^-1	0.85468	0.236225
1	0.85501	0.780598

5. (1%) 請討論你認為哪個attribute對結果影響最大？

觀察標準化後的logistic regression model的w參數,可發現train_X的第四欄（capital gain）之參數影響最大,如果只用capital gain是否>5000來判斷,就有 Private Score: 0.739977; Public Score :0.80405的表現,幾乎跟generative model表現一樣,因此我認為capital gain對結果影響最大。