Homework 2 Report - Income Prediction

學號：b04501095 系級：土木三 姓名：黃平瑋

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

	public score (accuracy)	private score (accuracy)
generative model	0.76474	0.76280
logistic regression	0.86351	0.85898

由上表所示logistic regression不論是在public還是private的資料上，表現都比generative model好很多

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

在觀察過資料的123個feature後發現“education”和“education-num”為類似的特徵，故刪除“education”只取“education-num”，並加入剩餘資料的二次項，再將所有資料經標準化(standardization)處理。
訓練的model運用logistic regression
loss function = cross entropy
optimizer = Adam
batch size = 32
epoch = 15000

training set accuracy = **0.862578** %
public accuracy = **0.86351** %
private accuracy = **0.85898** %

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)

標準化的方式我採用standardization，使每個feature變成平均為0，標準差為1的分佈，
平均和標準差都是從training set抽取出來，再同時套用到training set & testing set 作scaling
底下是我作data processing所使用的code

```python
def standardize(training_set, testing_set):
    mu = np.mean(training_set, axis = 0, dtype=np.float64)
    sigma = np.std(training_set, axis = 0, dtype=np.float64)
    mu_1 = np.tile(mu, (training_set.shape[0], 1))
    sigma_1 = np.tile(sigma, (training_set.shape[0], 1))
    mu_2 = np.tile(mu, (testing_set.shape[0], 1))
    sigma_2 = np.tile(sigma, (testing_set.shape[0], 1))
    training_set = (training_set - mu_1) / sigma_1
    testing_set = (testing_set - mu_2) / sigma_2
    return training_set, testing_set
```

下表是針對標準化對準確率的分析

	training set score (accuracy)	public score (accuracy)	private score (accuracy)
with standardization	0.862578	0.86351	0.85898
without standardization	0.771874	0.71842	0.71330

很顯然的資料經過標準化處理後，能更忠實的呈現各feature的特性，也能達到較高的準確率，原本資料有連續的feature如age,capital_gain，也有用one-hot encoding的離散資料，若沒有經過標準化的處理，這兩種資料型態差距太大，將無法產生教好的model

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)

在原本的loss function後面加上λw^2，取梯度後可以得到

$\nabla L(\boldsymbol{\theta}) = [\ \sigma\ (wx+b) - y\ ]x + \lambda w$

parameter λ	training set score (accuracy)	public score (accuracy)	private score (accuracy)
0.1	0.862578	0.86351	0.85898
1	0.860429	0.85982	0.85677
10	0.858825	0.85859	0.85714
100	0.856607	0.85773	0.85517

可能是資料有經過標準化的處理，如果 λ太大，使model過於注重參數的圓滑程度，準確度反而都沒有 λ小來的好，以本次實驗來說 λ = 0.1表現最好。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

我將所有model所有的weight印出來後，發現年齡的weight是裡面中最大的，有3.978，其他的參數weight則大多數小於1。這顯示出收入和年齡是有高度正相關的，而我之後也加入了年齡的高次項，也讓預測的準確度提昇了一些。