學號：B06705057　系級：資工二　姓名：黃資翔

1. 請比較你實作的generative model、logistic regression 的準確率，何者較佳？
(各個特徵皆只有一次項)

	Public score	Private score
Logistic regression	0.85171	0.85345
Generative	0.81965	0.81722

Logistic regression 表現較好，因為 generative model 中我們假設資料為 Gaussian distribution

2. 請說明你實作的best model，其訓練方式和準確率為何？
Best model 也是用 logistic regression，並將離散的特徵換成連續，如 education 換成 education_num ，國家換成國民平均收入。將這些連續項（共 7 個），加上二次項與三次項進行 training

Public score	Private score
0.85921	0.85837

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
在相同 model （X_train 中特徵只取一次項的logistic model）、相同迭代次數（10000 次）、相同找極值方法的情況下，僅有是否標準化的差別，結果如下：

	Public score	Private score
Normalization	0.85171	0.85345
non - normalization	0.79287	0.79240

標準化的影響：

1. 標準化前有些值太大，通過 sigmoid function 會 overflow，必須使用更小的 learning rate 導致收斂過慢
2. 標準化前，圖形為超橢圓形，每個特徵迭代的次數不同，較難實作
3. 標準化前可能迭代很多次效果也不會比標準化好，可能是即使當斜率很小時，仍在最佳解的遠處，而小數的位數是有限的，當斜率為數值 0 時則停止更新，但此時仍不是最佳解

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。

	Training score	Public score	Private score
Lambda = 100000	0.85326	0.85442	0.85439
Lambda = 10000	0.85639	0.85601	0.85726
Lambda = 1000	0.85627	0.85712	0.85726
Lambda = 100	0.85612	0.85700	0.85751
Lambda = 10	0.85612	0.85700	0.85751
Lambda = 0	0.85612	0.85700	0.85751

由 Training score 得知準確度先降後升，表示實際上我們並沒有找到 loss 最低的點，有沒有做 regularization 效果似乎不大，反而可能會有反效果。這個原因可能是我們並沒有找到最低點（並沒有 overfit training data），不需要再額外 regularize 也不會太差。

5. 請討論你認為哪個attribute 對結果影響最大？
我們將一個個 feature 刪除看結果

	Training score	Public score	Private score
Original data	0.85612	0.85700	0.85751
Delete age	0.85182	0.85147	0.85382
Delete fnlwgt	0.85664	0.85515	0.85628
Delete capital gain	0.75430	0.75061	0.74769
Delete capital loss	0.85433	0.85565	0.85198
Delete hour per week	0.85627	0.85454	0.85689
Delete country	0.85547	0.85651	0.85738
Delete education	0.84985	0.84936	0.85087
Delete sex	0.85654	0.85626	0.85824
Delete work class	0.85510	0.85712	0.85468
Delete marital status	0.85528	0.85638	0.85640
Delete occupation	0.85305	0.85331	0.85321
Delete relationship	0.85636	0.85700	0.85751
Delete race	0.84715	0.84803	0.84879

由結果知，capital gain 對準確度影響最大，其次是 education 與 race