學號:R07942115 系級:電信碩一 姓名:謝硯澤

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

	Public score	Private score
generative model	0.82800	0.82152
logistic regression	0.85171	0.85345

logistic regression 在 public 及 private score 都好很多。

2. 請說明你實作的 best model，其訓練方式和準確率為何?

我使用 scikit-learn 裡面的套件 GradientBoostingClassifier，

訓練方式：根據 release 的檔案(X_train)，裡面所有的特徵全部拿去做 training，除了固定 random seed，我還設定 n_estimators=300，除此之外，我並沒有再做其他的參數設定。

準確率如下表：

	Public score	Private score
GradientBoostingClassifier	0.87690	0.87483

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

使用 logistic regression、全部的特徵，iteration 設為 10000，並比較有無加 normalization 的差異。

	lr	Training loss	Public score	Private score
With normalization	0.5	0.31579	0.85171	0.85345
Without normalization	10^{-10}	0.5388	0.79594	0.79191

沒有加入特徵標準化的話，learning rate 要設得非常小，才有辦法 train 起來，而表現上也比較差，可能的原因是我是抽取全部的 feature，每個 feature 的值差異很大，使得 sigmoid 的值很容易被這個 feature 影響，造成訓練上的困難。

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。

加入正規化後，基本上沒有的影響～可能是因為我的 model 並沒有過擬合的問題，所以加入 regularization 後的影響並不明顯。

5. 請討論你認為哪個 attribute 對結果影響最大?

使用 logistic regression 來討論選擇不同 attribute 對結果的影響 ：

	All feature	w/o **fnlwgt**	w/o age	w/o **hours_per_week**
Public score	0.85171	0.85233	0.85171	0.85307
Private score	0.85345	0.85136	0.85173	0.84940

	w/o capital gain	w/o capital loss
Public score	0.84066	0.85122
Private score	0.83417	0.84780

我總共做了六組實驗，討論刪除不同 attribute 對結果的影響，出乎意料的，我原本以為刪除 age 會影響最大（主觀認為年紀越大，收入會越高），但看起來 capital gain 的影響才是最大的，根據上表，當拿掉 capital gain 時的分數下降很多！變化最為明顯～