學號：B05901068 系級：電機三 姓名：蕭如芸

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

	Generative model	Logistic regression
Public	0.84619	0.85884
Private	0.84105	0.85861

根據 kaggle 上 public 和 private 的分數，logistic regression 的準確率較高。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

我的 best model 使用 logistic regression 訓練，有使用 L1 regularization ($\lambda = 0.001$)。
先測試每個 feature 的重要性，方法是對於 training data，每次拿掉一個 feature，用剩下的 feature 在 generative model 上訓練，觀察在 training data 上的正確率。
我的 model 使用 X_train、取出的重要 feature 的 2 次方到 5 次方、log(1+X_train)，以上全部串在一起，再做 standardization。
在 kaggle 上的準確率為：public 0.85884, private 0.85861

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。

實作 standardization：$z = \frac{x-\mu}{\sigma}$
比較是否有做 normalization 的準確率（用 kaggle 上 public 和 private 的平均）

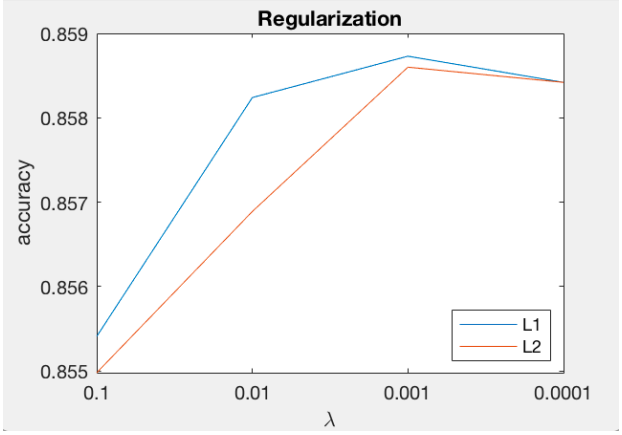
	Generative model	Logistic regression
沒有 normalization	0.84368	0.76377
有 normalization	0.84362	0.85873

是否有做 feature normalization 對 generative model 幾乎沒有影響；但對 logistic regression 影響很大，準確率相差接近 10%。

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。

比較 L1 regularization 和 L2 regularization 在不同 λ 值的準確率（用 kaggle 上 public 和 private 的平均）

λ	L1 Regularization	L2 Regularization
0.1	0.85541	0.85498
0.01	0.85824	0.85689
0.001	0.85873	0.85860
0.0001	0.85842	0.85842



從上圖可見，不論使用 L1 或 L2 regularization，準確率對不同 λ 值的趨勢大致上是相同的，$\lambda = 0.001$時的準確率最高。而 L1 regularization 的效果比 L2 好，原因可能是我的 model 包含一次項到五次項，L1 regularization 會得到比較 sparse 的解，可以避免 overfitting。

5. 請討論你認為哪個 attribute 對結果影響最大？

用 generative model 和整理過的 feature, X_train 進行分析。先拿完整的 training data，用 generative model 訓練，得到在 training data 上的正確率為 0.84245。接著每次拿掉一個 feature，用剩下的 105 個 feature 訓練，重複 106 次，觀察在 training data 上的正確率，若正確率明顯下降就表示拿掉的是重要的 feature。

實驗後發現，大部分的 feature 拿掉後，training accuracy 變化不大，只有幾項會明顯下降，以下是下降最多的五項 feature 及所對應拿掉後的 training accuracy：

capital_gain	0.83505
hours_per_week	0.84027
capital_loss	0.84048
age	0.84061
fnlwgt	0.84174

根據以上實驗結果，我認為 “capital_gain” 對結果影響最大，因為拿掉這個 feature 後，training accuracy 下降最多。