學號：R06921037 系級： 電機碩一 姓名：許哲瑋

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

答：

	Generative model	**Logistic regerssion**
Training(十次平均) (validation accuracy)	**84.3366%**	**85.1265%**
Testing(Public)	**84.287%**	**85.466%**

實作的 Generative model 的準確率無論是在 training set 和 testing set 上的表現都稍微差了一些。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

實作出來的 best model 即為 logistic regression 的 model，把婚姻的資料(marital status)的七種 marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse 分成 married 和 Not_In_Marriage. 將 native-country 分成 South-America，US_Canada，Europe_1(西歐北歐和英國)，Europe_2(東歐)，South-East-Asia，其餘保持原樣。Job 分成勞力/非勞力/服務業。訓練的準確率為 85.466%。

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。

答：

	Generative model	**Logistic regerssion**
Training(十次平均) (validation accuracy)	**84.4595%**	**79.2645%**
Testing(Public)	**84.336%**	**75.2381%**

當未標準化時，generative model 的準確率在 valiation accuracy 和 testing accuracy 都稍微的提高了，但是實際上的影響不大，若以兩個 class(收入大於 50K 或小於 50K)來說，假設兩個 class 都從 gaussian distribution 中選出了最大機率的 instance，那麼將資料特徵標準化對於辨別出這兩個高斯分布的機率影響並不大，換句話說，並不會讓兩者分得比較開/不開。

而標準化對於 logistic regression 的影響就比較巨大，若未標準化，丟進 sigmoid 之後的變形會更劇烈，預測也更不準。

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。

答：

Generative model						
	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
Training (十次平均) (validation accuracy)	**84.9842%**	**85.1194%**	**83.4528%**	**76.2663%**	**82.4864%**	**82.2234%**

做完 regularization 後發現，在λ 等於 0.01~10 中間似乎有準確率下降的影響，但更大後似乎就固定一個值不變。

5.請討論你認為哪個 attribute 對結果影響最大？

原本 85.38%

考慮:

拿掉國籍 85.23%

拿掉人種 85.17%

拿掉家庭育兒狀態 85.21 %

拿掉工作型態(不同的工作) 84.61%

拿掉婚姻狀態 85.11%

拿掉教育程度 84.65%

拿掉工作類別(是不是自僱者)85.22%

拿掉聯邦或地方政府 ：85.23%

拿掉 hour_per_week : 85.12%

拿掉 capital_gain capital_loss：83.61%

拿掉 sex ：85.26%

拿掉 fnlwgt : 85.21%

拿掉 Age: 85.11%

就上述討論，capital_gain 和 capital_loss 似乎影響最大!