學號：B05203050 系級：化學三 姓名：陳品翰

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

	private	public
generative model	0.84215	0.84619
logistic regression	0.85136	0.85307

logistic regression 較佳，應是由於資料的分布與 Gaussian 相去太遠，且資料夠多，故 generative model 沒有優勢。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

A. 資料前處理：

由於將 0 和 1 進行 normalize 只會變成另外兩個不同的數，我認為意義不大。故只將所有非 one hot encoding 的資料進行 normalize。之後進行 logistic regression，發現 fnlwgt 項的權重遠小於其他，故將此項刪除，以簡化模型，避免 overfit。另外將原本只算一個 attribute 的 gender 拆為兩項，意即在每組資料的最後加上與原 gender 項相反的數。

B. 模型架構

於 best model 中採用 logistic regression，以 keras 套件協助訓練，加上一層有 10 個 neuron 的 hidden layer，並在輸入層加入 regularizers.l2(0.001)。輸入層到 hidden layer 的 activation function 為 relu，hidden layer 到輸出層的 activation function 為 sigmoid，optimizer 為 RMSProp (實作發現比 Adam 好)。batch size 為 100，epochs 為 100。

C. 結果

	private	public
logistic regression	0.85136	0.85307
best model	0.85775	0.85577

可見 best model 的 accuracy 有顯著提升，variance 也很小。

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

generative model	private	public
without normalization	0.83687	0.84287
with normalization	0.84215	0.84619

logistic regression	private	public
without normalization	0.78344	0.78181
with normalization	0.85136	0.85307

logistic regression normalize 後，準確率有顯著提升，而 generative 的提升程度較小。由於 normalize 後，對於 gaussian 的分布情況不會有影響，只是將平均值改為 0，標準差改為 1。

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。

logistic regression	private	public
without regularization	0.85136	0.85307
with regularization	0.85210	0.85380

兩者 variance 都很小，無太大差距，但加上 regularization 後，準確率有所提升（$\lambda = 0.0001$）。

5. 請討論你認為哪個 attribute 對結果影響最大？

將 logistic regression 的所有 weight print 出來，發現絕對值最大的項目是 capital gain，為 1.67。而非連續的項目中絕對值平均最大的是 education，為 1.03，故此二者影響最大。