學號：B04901147 系級：電機四 姓名：黃健祐

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

model type	public	private
generative	0.84570	0.84080
logistic	0.84361	0.84105

在 public set 上，generative model 表現較佳，private set 上則被 logistic model 稍微超越。就整體而言，兩者表現差不多，但 generative model 略勝一籌。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

best model 捨棄了國籍以及 fnlwgt 的資訊，因此 input 為 64 維的向量，並且每個維度都有進行 normalization。訓練時則是使用 sklearn 的 gradient boosting。最後準確率在 public set 上為 0.86744、private set 上則是 0.86389，兩者皆通過了 strong baseline。

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

model type	public	private
generative – with normalization	0.84570	0.84080
generative – without normalization	0.84668	0.84092
logistic – with normalization	0.84361	0.84105
logistic – without normalization	0.78660	0.78651

由上表可以發現，有無 feature normalization 對 logistic model 影響甚大，但 generative model 則沒有明顯的變化。在 logistic model 的訓練及測試過程中，數

值較大的 feature 可能會主導 gradient 的數值及方向，使得某些參數沒辦法獲得很好的調整；generative model 則是因為直接透過公式計算，因此較不受影響。

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。

λ	train	dev	public	private
0	0.84406	0.84091	0.84373	0.84031
0.0001	0.84402	0.84173	0.84361	0.84031
0.001	0.84352	0.84091	0.84398	0.84006
0.01	0.83752	0.83804	0.84017	0.83552
0.1	0.80742	0.80610	0.81117	0.80714
1	0.76370	0.75573	0.76977	0.76513

由上表可以得知，適度的 regularization 對於模型的表現是有幫助的，但過大的λ則會使得 model 為了平衡每個 feature 之間的權重而犧牲了準確率。

5. 請討論你認為哪個 attribute 對結果影響最大？

觀察 logistic model 的參數可以發現，capital gain 的權重最大。若只根據 capital gain 來預測，也可以達到大約 0.6 至 0.7 的準確率，因此我推測 capital gain 的影響是最大的。