學號：B04611041 系級： 工海三 姓名：簡暐晉

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

答：

generative model 的 public 和 private 的分數分別為 84.533%，84.191%，logistic regression 的 public 和 private 的分數為 85.749%和 85.530%，結果為 logistic 的結果比較好，推測應該是因為 generative model 假定 data 為 Gaussian distribution，但可能並非所有的 feature 分布都是 Gaussian distribution，因次結果可能受影響。而因 logistic regression 並沒有預設資料是如何分布的，因此結果比較好。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我使用 scikit-learn 中的 GradientBoostingClassifier，將 n_estimator 設為 400，max_depth=3，並新增 fnlwgt, age，hours_per_week 的 2 次方，fnlwgt 的三次方為 feature，之後將所有的 feature 做 normalization，得到的結果為 Public 87.825%，private: 87.495%。

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

答：

Generative		
Normalization	public	private
yes	0.84582	0.84252
no	0.84533	0.84191

logistic regression		
Normalization	public	private
yes	0.85749	0.8553
no	0.79287	0.7929

在 generative model 中有沒有做 Normalization 的結果並沒有差很多，推測試因為做完 normalization 後因為分布不變，因此對結果沒有什麼影響。但對於 logistic regression 的結果卻影響很大，應該是因為原本的資料中有連續也有離散的資料，而連續資料的分布因為範圍很廣，直接 train 的話對於 weight 會有比較大的誤差。

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：

λ	public	private
0.1	0.85749	0.8553
10	0.85749	0.8553
100	0.85749	0.8553
1000	0.85749	0.8553
1500	0.85749	0.85517
2000	0.85761	0.8548
2500	0.85773	0.8548
2500	0.85761	0.8548

因為 w 非常小(0.01 以下)，因此 regularization 可能對結果不會產生太大的影響。

5.請討論你認為哪個 attribute 對結果影響最大？

在做 logistic regression 時我把 fnlwgt,age,hours_per_week,各取了 1~4 次方，另外新增一個 feature 為 capital_gain-captial_loss 的值，再加上原本的 feature，得到的效果最好，因此推測這幾個 feature 對 logistic regression 的影響比較大的，後來我又一一將 fnlwgt,age,hours_per_week,capital_gain-captial_loss 這些 feature 拿掉做實驗，發現拿掉 capital_gain 時的結果最差，因此在這些 feature 之中可能是 capital_gain 對我影響最大。