學號：B03901145 系級： 電機四 姓名：郭恆成

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

accuracy	Public set	Private set
Generative model	0.84533	0.84215
Logistic regression	0.86732	0.86549

有 kaggle 上的 Leaderboard 的成績可以發現不管在 public、private 上的表現都是 Logistic regression 上表現較佳。
推測是因為 generative model 實做上是假定 data 是 Gaussian distribution，去計算機率密度，而 Logistic regression 則沒有限制，所以 Logistic regression 若 feature 取的夠好，就能更好的去預測 fit data 的機率分佈。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

accuracy	Public set	Private set
XGBClassifier	0.86732	0.86549

我的 best model 是用 xgboost 裡的 XGBClassifier，XGBClassifier 是 gradient boosting 算法的優化。

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。

accuracy	Public set	Private set
Generative model	0.84508	0.84227
Generative model (normalize)	0.84533	0.84215
Logistic regression	0.23768	0.23476
Logistic regression (normalize)	0.86056	0.85579
XGBClassifier	0.86732	0.86549
XGBClassifier (normalize)	0.86732	0.86549

- Generative model : 因為 normalize 會讓 Gaussian 的(mean , std) =(1, 0)，但因為同時對 test data 也做 normalize，對彼此相對的機率分佈影響不大，所以兩個 model 的準確率差不多。
- Logistic regression: 猜測是我 feature 取了部分 data 的五次項以及 log 項，所以是否 normalize 對 feature 的 scale 影響極大，而實作上發現沒有 normalize 訓練不起來，準確率會週期性的震盪。
- XGBClassifer: 兩者的成績相同。

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。

Lambda	Public	Private
0	0.86056	0.85579
0.01	0.86056	0.85579
0.1	0.86056	0.85579
1	0.86068	0.85591
10	0.86191	0.85665
100	0.85970	0.85616

可以看到當 lambda 小於 1 時 regularization 數字太小對於準確率沒有影響，而 lambda 等於 10 時會得到最佳的準確率。lambda 在繼續增加準確率就會下降。

5.請討論你認為哪個 attribute 對結果影響最大？

把幾個數據拿掉得到的準確率如下:

	Public	Private
all	0.86056	0.85579
age	0.85921	0.85272
fnlwgt	0.85859	0.85554
sex	0.86117	0.85689
capital gain	0.85970	0.85603
capital loss	0.85663	0.85296

由此可知，在我設計的 model 中 age 對準確率的影響最大。