學號：R05631027 系級： 生機碩二 姓名：楊皓文

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
答：
generative model: 先利用 Gaussian distribution 預估 mu1, mu2, shared sigma 後，可以得到 weight 和 bias，最後計算 likelihood 來分類。
logistic regression: 本次實驗 optimizer 為 Adagrad，learning rate 為 0.5，batch size 為 320，epoch 為 2000。利用 sigmoid function 將分成 0~1，再計算 loss function 調整 weight 和 bias。
在兩者皆只有用 X_train 的 106 個 attributes 情況下，我們可以發現 logistic regression 的準確率較佳，如下表所示。

| | generative model | logistic regression |
|---|---|---|
| 10-fold validation accuracy | 0.835074 | 0.847973 |

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
我的 best model 為 logistic regression，使用 X_train 的 106 個 attributes 加上 age、fnlwgt、capital_gain、cappital_loss、hours_per_week 各別的的平方、三次方 和四次方項，總共 121 個 attributes。
optimizer 為 Adagrad，learning rate 為 0.5，batch size 為 320，epoch 為 2000。最後的十折交叉驗證準確率為 0.855037，平均 loss 為 0.300710。

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：
由於 generative model 和 logistic regression 都需要經過 sigmoid function，因此若沒有進行特徵標準化(feature normalization)的動作，很容易產生 overflow，造成整個模型壞掉或準確率降低，因此在這兩個建立模型的方法上，做特徵標準化是必須的。

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：
本次實驗使用 X_train 的 106 個 attributes ，optimizer 為 Adagrad，learning rate 為 0.5，batch size 為 320，epoch 為 1000。
由下表可以看出 regulariztion 的影響並不明顯，準確率都相同，只有 c 為 10 會有稍微較好的 loss 表現。

| Regularization constant | c = 0.0 | c = 10.0 | c = 1.0 | c = 0.1 | c = 0.01 |
|---|---|---|---|---|---|
| 10-fold validation accuracy | 0.847666 | 0.847666 | 0.847666 | 0.847666 | 0.847666 |
| epoch avg loss | 0.314948 | 0.314902 | 0.314945 | 0.314948 | 0.314948 |

5.請討論你認為哪個 attribute 對結果影響最大？

答：

我的實驗是先輸入一組一次方的全部 attributes，在加上任一個二次方 attribute 項後，我們可以發現加上二次方 age 值可以讓 loss 快速下降，比加起其他二次方項有更好的表現。另一方面以實際日常生活來看，年齡和收入的確有著一定的關聯性，通常年齡較高收入也會較多，因此我認為 age 這個 attribute 對結果影響最大。