學號：B04902105 系級： 資工三 姓名：戴培倫

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

[generative model]
自己train出來的準確率只有0.75919，而public score只有0.76523。嘗試加上幾個參數的二次方(前五項)也是同樣的結果，通過sigmoid function後的值大部份都小於0.3。

[logistic regression]
自己train的準確率大約0.8574, public score最佳為0.86019，比generative model 好非常多。

2.請說明你實作的best model，其訓練方式和準確率為何？
答：
Best model使用的是logistic regression，取用了所有feature的一次項、前五項feature的2~8次項。取完所有feature後，與testing data一起做normalization。
learning rate = 0.1，epochs = 2000，並使用Adagrad更新weight和bias。
自行取10%的training data做validation算出來的準確率為0.8574，kaggle public score為0.86019，kaggle private score為0.85677。

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：

沒有normalize時因為我的best_model取用了高次方，在計算sigmoid function時會overflow，因此我只用到二次項來做測試。以下為kaggle public score以及自行切10% training data的validation score。

[沒有normalize]
Public score: 0.80294
Validation: 0.78969
[有normalize]
Public score: 0.84447
Validation: 0.85257

推測原因： data中的fnlwgt以及capital_gain, capital_loss數值從數千至數十萬，而其他feature大部份值為0或1，數值範圍過大，因此normalize對於train model很有幫助。

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	λ = 0	λ = 0.01	λ = 0.1	λ = 1
準確率	0.86019	0.85909	0.85847	0.85884

準確率只相差約0.2%，相差很小，推估是因為算出來的weight很小，故正規化對模型的準確率幫助不大。

5.請討論你認為哪個attribute對結果影響最大？

我從一次方的model中發現前五項的weight較大，認為他們對結果影響較大，因此取出他們做更高次的分析。發現其實每一項對準確率都有幫助，因此我將每一項分別加到6次方做validation，發現第一項，也就是age在validation中表現最好，因此我認為age對結果影響最大。