Homework 2 Report - Income Prediction

學號：b04901009　系級：電機三　姓名：林孟瑾

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

	Private score	Public score
generative model	0.83638	0.83906
logistic regression	0.85665	0.85933

generative model 較為不準確，private score 與 public score 都比 logistic regression 差。由於 generative model 只是以機率的方式預測，無法描述 data 中較複雜的行為，所以準確率較低。而 logistic regression 產生的模型為多維的 weight，較能接近 data 的複雜分佈，所以準確率較佳。

2.(1%) 請說明你實作的 best model，其訓練方式和準確率為何?

我的 best model 是運用 logistic regression。選擇適當的 features 之後，再加上 continuous 項的 1.2 次方和 1.4 次方，如此較低的次方接近 linear，較不會有 overfit 的情況。Private score 為 0.85665，public score 為 0.85933。

3.(1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

	No normalization		normalization	
	Private score	Public score	Private score	Public score
generative model	0.83859	0.84115	0.83638	0.83906
logistic regression	0.79437	0.80122	0.85665	0.85933

Logistic regression 中，normalization 會讓準確率提升，因為不同 feature 間，在做 gradient decent 時不會過於偏向某些 feature 而影響結果。而 generative model 中，normalization 對準確率的影響較不明顯，可能是因為機率模型是依據每個 feature 出現的機率分佈，feature 本身的 scaling 並不重要。

4.(1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

	Private score	Public score
regularization	0.85665	0.85933
No regularization	0.78761	0.78955

Regularization 讓準確率提升，因為 regularization 能夠使得模型較不 overfit，讓 model 較為平滑，進而提升準確率。

5.(1%) 請討論你認為哪個 attribute 對結果影響最大?

原本的 logistic regression 的 private score 為 0.85665，public score 為 0.85933。

	Private score	Public score
拿掉 age	0.85186	0.85626
拿掉 fnlwgt	0.85603	0.85872
拿掉 capital_gain	0.84326	0.84226
拿掉 capital_loss	0.85259	0.85589
拿掉 hours_per_week	0.85480	0.85663

將 data 中的某些 continuous 項去除後然後用 logistic regression 訓練 model，發現 private score 和 public score 下降最多的是“拿掉 capital_gain”，表示此項對於結果影響最大。