Homework 2 Report - Income Prediction

學號：b05901041　系級：電機二　姓名：蘇家軒

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

自己實作的部分，generative model 較佳。推測原因是自己的 logistic regression 程式有部分的參數沒有調整好造成結果不佳;而 generative model 因為實作相對容易而沒有太多差錯。

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

我去掉 fnlwgt 這個 attribute 後使用 sklearn 作 logistic regression，在 public 上取得 85.712%的正確率;private 則獲得 84.8%的正確率。另外試著使用 keras，用 neuron network 做 logistic regression，public set 的正確率為 85.5%，private 正確率為 85.4%。

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)

在 logistic regression 的部分，我認為 normalization 是相當必要的，因為若沒做 normalization，常常一不小心就造成在算 sigmoid 時容易出現 math range error(sigmoid input 過大)。

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)

我試著使用 sklearn 套件調整 regulation 參數去做，結果在 private 和 public set 的準確率幾乎一模一樣(差不到 1%)，我認為正規化的有無並無太大影響。

5.(1%) 請討論你認為哪個 attribute 對結果影響最大？

在 training set 中，Income>50k 的人裡面，有高達 85%的人有配偶，而如果只依據有無配偶來決定 income 是否大於 50k(沒有配偶猜不大於 50k，有的話猜大於 50k)也有高達 70%的正確率，因此我認為此 attribute 對結果有相當影響。