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poblic上取得85.712%的正確率。

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。

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)

基本上對於準確率幾乎沒有太大的影響，不過也有可能是我logistic regression本身做爛了，絕大部分的預測結果都是0的原因。

我也試著用sklearn套件調整regulation參數去做，結果public set的準確率一模一樣。

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。