Homework 2 Report - Income Prediction

學號：r06521705 系級：土木系營管組碩一 姓名：陳思愷

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

比較過後認為 logistic regression 的準確率較佳較佳，可能是因為 generative model 對於資料的假設並非每個都符合實際情況(例如: feature 之間彼此要獨立這件事就不符合本題之情況)，且資料量不算少，所以主要吃資料量來決定準確率的 logistic regression 會較為準確。

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

Best 的檔案當中，引入了 keras 的套件，但由於了解甚少，為何結果會變得比較好還是有點不知道原因，但可能是因為套件的各部分(ex, optimizer , validation 等等)都寫得比自己寫的還要更加完善許多。
Optimizer 選用的是 adam
Loss 則是以 cross entropy 來呈現

Public: 0.86142
Private: 0.85837

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)

從Logistic model來看，在未做標準化的情況下，Public跟Private score都只有將近0.78，但實作feature normalization後，可以上升到0.86附近，可能是因為feature當中有些會`大到3、40000但有些卻只有0、1，feature的scale相差太大，可能會影響準確度

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)

λ	1	10	100	1000
Public	0.85365	0.85468	0.85321	0.82856
Private	0.84952	0.85172	0.84720	0.82362

當 lamda 設在 10 的時候會有較佳的準確率

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

我認為 9 個 workclass、7 個 marital、15 個 occupation、2 個 sex(from 不是 csv 的 training data)
再加上 fnlwgt、education、capital_gain、capital_loss、hours_per(from csv 的 training data)
是對結果影響較大的 attribute。