Homework 2 Report - Income Prediction

學號：b04705043　系級：資管三　姓名：張凱庭

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

	Public score	Private score
Logistic	0.85652	0.85823
genrative	0.84619	0.83982

兩個 model 使用一樣經過 normalize 後的資料，根據 kaggle 上的 socore 進行比較，

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

使用 logistic regression 得到在 kaggle 最佳的結果

Logistic	0.85652	0.85823

對 continus variable 進行 normalize，訓練時將 continus variable 加入 2~5 次方項，使結果更 fit。訓練中使用 adagrad，訓練次數=5000，learning-rate = 0.05

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)

	Public score	Private score
Logistic – normal	0.85652	0.85823
genrative -normal	0.84619	0.83982
Logistic – unnormal	0.2	
genrative -unnormal	0.63808	0.63628

可以看出不論使用哪種做法，標準化的資料訓練後的準確率有大幅提升，我想原因是這次的資料大多都是 dummy varible 的形式，只有幾個是 continus variable，訓練起來每個 feture 的 scale 差異過大使得準確率不高，將 continus variable 進行 normalize 後訓練起來每個 feture 的 scale 都差不多，得到的訓練結果準確率較佳

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)

Lambda	Public score	Private score
0.1	0.85652	0.85823

genrative -normal	0.84619	0.83982
Logistic – unnormal	0.2	
genrative -unnormal	0.63808	0.63628

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？