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，logistic 的表現較佳，猜測可能是因為，logistic 會試著找出給定 x，y 出現的機率 p(y|x)，generative 則會試著找出 p(x,y)，在分類任務上，屬於判別模型的 logistic regression 表現較 generative 佳。

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 在 kaggleg 上得到的分數：

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	0.23525	0.23719
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。其中還可以看到 logistic 時若沒經過正規化，在一樣的 learning rate 以及訓練次數，反而發散了，由此可知 feature normalization 在此預測中十分重要

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)

Lambda	Public score	Private score

0.001	0.85823	0.85652
0.01	0.85823	0.85652
0.1	0.85847	0.85628
1	0.85810	0.85616
10	0.85651	0.85542

經由數據可以看出正規劃並沒有顯著的影響，應該是因為正規化是用來解決 orverfit 的問題，而原先的 model 可能並不複雜，所以沒有產生 overfit，因此正規化並沒有太大的影響

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

我認為 continus variable 都對結果影響很大，其中又以 age 這個項目影響特別大。訓練時若將 age 移除 ，validation test 結果從原先的 0.84 掉到 0.78。而移除其他項目時結果變成 0.82 左右，所以認定 age 對結果影響最大