Homework 2 Report - Income Prediction

學號：b04501127 系級：土木三 姓名：凌于凱

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

   All feature : generative model: public score(0.84545)，private score(0.84166)

   logistic regression: public score(0.85724)，private score(0.84829)

   logistic regression 表現較好，可能是因為有些 feature 值並不是 binary，使得我實做 generative model 都是用高斯分佈下會比較不合適。

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

   Feature: all feature + log(age, captain gain, captain loss, hour per week + 1)

   訓練方法：sgd LogisticRegression (lr = 0.001, epoch = 10000, batchsize = 32)

   Accuratcy：public score(0.86007)，private score(0.85677)

   （原本是使用 sklearn LogisticRegression，結果 public score 和用 sgd 一樣，但 private score 為 0.85566 較低，可能是用 sgd 會造成 loss 震盪到較好的值，所以表現才必較好）

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)

   all feature logistic regression(lr = 0.001, epoch = 10000, batchsize = 32)

   沒有 normalize：loss(5.8560)，public score(0.79130)，private score(0.79852)

   有 normalize：loss(0.3158)，public score(0.85687)，private score(0.84805)

   因為不同的 feature 數值相差很大，故如果沒有 normalize 會造成 loss function 無法到達最低的值，所以有 normalize 的準確性會較好。

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)

   Feature: all feature + log(age, captain gain, captain loss, hour per week + 1)

   訓練方法：sgd LogisticRegression (lr = 0.001, epoch = 1000, batchsize = 32)

lambda	0	0.1	0.01	0.001
Public score	0.85798	0.85798	0.85835	0.85835
Private score	0.85542	0.85542	0.85566	0.85505

   可看出準確性並沒有大幅度的改變，可能是因為模型還不夠複雜，沒有到 overfitting 的程度。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大?

feature	accuracy	feature	accuracy
All feature	0.85169	-relationship	0.85117
-age	0.85016	-race	0.85163
-workclass	0.85034	-sex	0.85114
-fnlwgt	0.85129	-captain_gain	0.83587
-education	0.84220	-capital_loss	0.84963
-marital_status	0.85166	-hours_per_week	0.84967
-occupation	0.84601	-native_country	0.85090

   利用 cv=10 cross valid split 算出把所有 feature 刪除其中一項 attribute 的正確率，推測 captain_gain 對結果影響最大。