Homework 2 Report - Income Prediction

學號：r06725041 系級：資管碩一 姓名：彭証鴻

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

Generative model: Public score: 0.78243 Private score : 0.78295

Logistic model: Public score: 0.85761 Private score: 0.84743

由結果可看出，logistic regression 的準確率是較好的，因為 generative model 是有事先對 data distribution 做假設，若實際上 data distribution 並不符合其假設，準確率不會比 discriminative model 好。若是在 data noise 多且 data 數量少時的情況下，可試試 Generative model，或許會有不錯的效果。

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

利用 scikitlearn 的 MLPClassifier 來實作，在 training 前有做 feature normalization，activation function 採用 sigmoid function，solver 設為 sgd(stochastic gradient descent)，透過 adaptive method 調整 learning rate，bach_size 設為 1。此外還有做 early stopping，而 hidden_layer 僅一層，layer 中 neuron 數目為 125。

準確率：Public score: 0.85749 Private score: 0.85296

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。

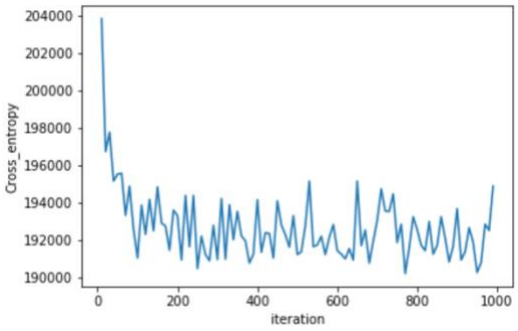
以下將以 logistic model 為例，比較 feature normalization 前後的差別：

固定參數：batch_size = 1, epoch = 1000, learing rate = 0.001

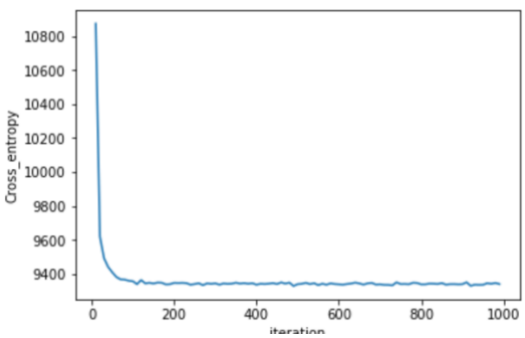
Feature Normalization 前：Public score : 0.79914、Private score : 0.79154

Feature Normalization 後: Public score : 0.85761 Private score : 0.84743

Normalization 前



Normalization 後



從上面兩張 cross entropy loss 圖可看出若沒有做 normalization，在 training 的過程中，收斂的速度慢且因 batch_size 設為 1 的關係，training 的過程非常不穩定，其結果也相當不好。

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 regularization(lambda = 0.0001)的 public score : 0.85466 private score : 0.84780
無 regularization 的 public score : 0.85761 private score : 0.84743

從 public、private 準確率的差來看，有 regularization 的 model，public 與 private 的準確率相差較小，由此可知有 regularization 的 model 較不會 overfitting，倘若參數調整得宜，其 private score 表現有機會比無 regularization 好。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

在嘗試挑掉不同的 attribute 的過程中，發現挑掉單個 attribute 對 model 的準確率沒有顯著的提升，但有發現挑掉”occupation”使得準確率有顯著的下降，所以在這次作業的 data 中，我們發現職業是對結果影響最大的 attribute。

實驗結果：
尚未挑掉任何的 feature 時，public score=0.85356 private score=0.84886
僅挑掉 occupation 後，public score=0.84729 private=0.84129
僅挑掉 age 後，public score=0.85380 private score = 0.84989
僅挑掉 marital status 後，public score=0.85319 private score = 0.84854
僅挑掉 education num 後，public score=0.85331 private score = 0.84866
僅挑掉 education 後，public score=0.85368 private score = 0.84854
僅挑掉 native 後，public score=0.85380 private score = 0.84915