Homework 2 Report - Income Prediction

學號：b04505026 系級：工海三 姓名：蔡仲閔

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

在我們 model 中，logistic 的準確率明顯較佳，這與老師上課提到，若是 data 足夠的前提下，使用 logistic 可以有較佳的結果有關。

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

在這次的作業中我試著使用 keras 實作的 neural network，但嘗試不同層數及方法後，準確率並不比 logistic regression 好太多，最佳的 model 使用的是 adam 作為 optimizer 並使用 binary_crossentropy 作為 loss。在 training 時可以輕易的到達 0.88 以上的準確率，然而 testing data 上則並不好，也嘗試使用 dropout，但最後仍無法達到 strong baseline。

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)

我們可以從下表中看到，在有 normalization 的 data 中有顯著較佳的結果，因為 normalization 可以避免部分 feature 因數值過大，比如說這次的 fnlwgt 或是 capital_gain，在經過 normalized 可以將所有 feature 壓縮到-1~1 之間，較平衡的去看待。而我也嘗試了 feature scaling 的作法 $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$，進而將 feature 壓縮在 0~1 之間，也有不錯的效果。

方法	Number of epoch	Private Score	Public Score
Normalized	2000	0.84682	0.85233
Without Normalized	2000	0.78872	0.79385

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)

在下表中我們可以看到 regulariztion 對於改善 training 的效果並不佳，一方面此方法本來就是希望獲得一個 smooth 的結果，而給予較大的 w 逞罰，並不保證在所有地方都適用。而我也試圖給與更多的 epoch，與對照組不同的，

regularization 的組別隨著 epoch 數增加有較佳的結果，但在 10 倍 epoch 數內仍不敵對照組的 performance。

方法	Number of epoch	Private Score	Public Score
Without reg	2000	0.84682	0.85233
With reg	2000	0.83650	0.84471
With reg	10000	0.84031	0.84668
With reg	20000	0.83920	0.84803

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

在討論哪個 attribute 影響最大時，我認為從最後 weight 的大小可以略知一二，我試著對 weight 作圖，並取大於一個平均加標準差(mu+sigma)的 weight 作討論，我們可以看到，其中對於收入正面影響最大的便是 capital_gain，也就是所謂的資本利得，這與我們所了解的現實社會相差不大，用錢賺錢仍是這個時代最快的賺錢方法。而對於收入有負面影響的則是前兩項，教育程度較低，以及第三、四項原國籍，我們看到來自海地以及哥倫比亞的人民普遍收入較低，這也與哥倫比亞毒梟橫行、海地天災人禍有關，因次也相當切合我們的觀察。