Homework 2 Report - Income Prediction

學號：R06521601 系級：土木碩 姓名：黃伯凱

1. (1%) 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
我的 generative model 準確率約為 80%左右，logistic model 約 85%左右，兩者差異滿大的。參考老師講義的說法是，雖然 generative 和 logistic 來自同一種模型，目的都是為了找到 w,b，但因為 generative 對訓練資料預先做了假設，它假設訓練資料符合機率模型這件事，這也導致兩方法會找到不同 w,b 的結果。也因為 generative 對訓練資料作了機率模型的假設，所以若訓練資料的差異相當大(如這次作業的 0 遠大於 1，0 約占 78%)，就很可能造成誤將 1 判斷為 0 的事件，也因此準確率較低。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？
我的 best model 使用了 keras 以及 tensorflow 兩套件，並且用 GPU 作運算，在經過資料處理後，我的特徵共有 39 項，所以我的第一層模型為 input 39，第二層為 output 1，並且使用 sigmoid 作為激活函數，learning rate 設為 0.001，epochs 設為 50，batch_size 設為 64，並且切分 10%訓練資料作為驗證資料，最後準確率大約為 86%左右。

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)
我第一次完成模型後，馬上進行我的第一次訓練，我發現我的準確率一直固定在 78%，一直上不去，調整了 learning_rate 也一樣還是收斂不了，卡關了一整天之後，突然想到上次作業的 feature scaling !!!馬上從床上跳起來對訓練資料作了 standernormalization 接著訓練就看到準確率開始起飛了，差點哭出來。我想是因為這次特徵有許多 0 或 1 以及數值的項目，但數值的項目相對 0 或 1 是很大的，它會對訓練結果造成太大的影響，所以肯定是要作 scaling。

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)
加入 regularization 之後對模型的準確率有稍微的提升，主要是因為，能夠避免曲線過於生硬，也就是所謂的 overfitting 或是 underfitting ，尤其是在若訓練的次數提高許多的情況下，更應該要做正規化，避免上述問題發生。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？
我覺得應該是 age 以及 capitalgain 和 capitalloss，我第一次訓練時，我是將兩項 capital 拿掉，因為我覺得這兩項資料好像太多 0，應該會對訓練結果造成誤差，但在我把這兩項加回去之後，準確率提升很多，也許這兩項攸關的是這個人的消費能力以及資本收益，所以是蠻關鍵的特徵，另外年齡通常也會和事業成就成正相關，所以這三項特徵我認為影響力最大。