Homework 2 Report - Income Prediction

學號：b05902052 系級：資工二 姓名：劉家維

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

Logistic Regression: 0.85503 / 0.84707

Generative Model: 0.74914 / 0.75224

Logistic regression 較佳。

相較於 generative model，logistic regression 屬於 discriminative model，兩者最大的不同是 discriminative model 學習如何判斷樣本資料（找出切開樣本資料的分界線），而 generative model 學習樣本的分佈。在 generative model 中，我們假設樣本是高斯分佈，但現實中的樣本通常不是這項分佈（這次一堆 attribute 資料都是 0 或 1，就非常不符合高斯分佈）。也許這就是為什麼 logistic regression 表現較佳的原因。

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

先剔除 fnlwgt 這項 attribute，再將資料標準化之後，使用

(1) logistic regression

(2) support vector classifier

(3) neural network

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 2048)	251904
dropout_1 (Dropout)	(None, 2048)	0
dense_2 (Dense)	(None, 256)	524544
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 1)	257

以上三個模型進行預測，看哪項預測結果佔多數，就選擇該預測結果。

準確率：(public / private) 0.86031 / 0.85186

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)

理論上對於 logistic regression 這種 linear model 來說，normalization 是屬於 linear transform，應該不會對準確率造成差異性。

但是不管是我自己手刻的 logistic regression 或者是用 sklearn 的套件，沒有做 normalization 都會訓練不起來，推測原因有以下幾點：

(1)

因為在運算過程中都會需要使用 exponential 函數，而這次各項資料的值的尺度差距非常大，若沒有 normalization 容易造成運算溢位或精度不足導致計算錯誤。

(2)

若沒有經過標準化，會難以找到較合適的 learning rate，導致不好收斂。

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)

λ = 1e-2, L2: **0.85724** / 0.84645, L1: 0.85712 / 0.84645

λ = 1e-1, L2: **0.85724** / 0.84645, L1: 0.85712 / 0.84645

λ = 1,   L2: **0.85724** / 0.84645, L1: 0.85761 / 0.84682

λ = 1e1, L2: 0.85712 / 0.84682, L1: **0.85773** / 0.84756

λ = 1e2, L2: 0.85663 / **0.84768**, L1: 0.85626 / **0.84768**

λ = 1e3, L2: 0.85024 / 0.84559, L1: 0.83353 / 0.82741

λ = 1e4, L2: 0.83452 / 0.82213, L1: 0.76474 / 0.76280

λ = 1e5, L2: 0.76400 / 0.76575, L1: 0.76474 / 0.76280

如想像中的一樣，λ 值太大會造成模型無法對已知資料做很好的擬合，太小則會造成過擬合。經過實驗後發現 L2 與 L1 的表現差不多，但是 L2 在 public 和 private 的表現落差較 L1 大，雖然這可能只是在 split test set 時造成的誤差。

5.(1%) 請討論你認為哪個 attribute 對結果影響最大？

capital_gain 和 age。

經過 logistic regression 訓練出來的參數，發現 capital_gain 這項係數約為 -2.37, age 這項係數約為 2.09，其他都介於 -0.6 ~ 0.6 之間，故我認為這兩項 attribute 對結果影響最大。