Homework 2 Report - Income Prediction

學號：b05902013 系級：資工二 姓名：吳宗翰

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

以下是我取相同feature使用logistic regression和generative model測出的結果： (在這裡的logistic regression是有tune過參數的，而generative model則是使用上課建議的把兩個Covarience matrix做加權)

model	Public Score	Private Score
logistic regression	0.85552	0.85210
generative model	0.84582	0.84363

結果是logistic regression略勝一籌，我想原因可能有以下幾點：

1. logistic regression中最關鍵的sigmoid function可以有效的把「表現過於突出」的資料給壓在$[0, 1]$這個區間內，因此他比較能對抗極端資料，找到適當的hyperplane去切開這個高維空間
2. 如同課堂上所提到的，在維度很高的狀況下，generative model的covarience matrix會導致varience變大，再加上generative難以做regularize才導致predict的時候差了些

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

有關於本次作業的best model，我用了以下方法訓練：

1. 資料preprocessing：我取了原本助教給的123維，另外還加上age, capital gain, capital loss, hours_per_week這四個參數的2到100次方上去(non-linear transform)。另外為了讓Gradient Descent容易些，我還有做feature scaling(Standardize)。
2. 訓練模型：logistic regression(使用sklearn套件)，大致都用現成default值，比較特別的是我使用了L1的Regularization，約束項$C = 3.9$。
3. 準確率如下： public score: 0.87567, private score: 0.87004

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)

以下是對我的`hw2_best_train.py`做了以下三種特徵標準化的結果：

標準化方法	數學公式	Public Score	Private Score
None	None	0.80294	0.79633
Rescaling	$x' = \frac{x - min(x)}{max(x) - min(x)}$	0.85749	0.84817
Standardization	$x' = \frac{x - \bar{x}}{\sigma}$	0.85712	0.84805

在這次的實作中，我觀察到了以下結果：

1. 在沒有做標準化直接train下去的結果其實蠻慘的，正確率大概只有0.8左右，符合上課所說的做feature scaling會讓梯度下降容易地些
2. 以實驗結果可以看到rescaling和standardization的performace其實差不多，因此應該是一個有做就好的東西

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)

以下是我改一下我的`hw2_best_train.py`用做了以下4種L2 regularization的結果：

C	Public Score	Private Score
0.0001	0.84668	0.84375
0.01	0.87432	0.87028
1	0.87555	0.87028
100	0.87616	0.87016
10000	0.85700	0.84805

在這次的實作中，我觀察到了以下結果：

1. 由於在`hw2_best_train.py`中的參數很多，如果$C = 0.0001$這麼小的話，確實可能會讓靈敏度過高而錯判許多的case
2. 由實驗結果發現regularization的penalty也不能開太大的，像是實驗結果顯示$C = 10000$反而讓accuracy下降

5. (1%) 請討論你認為哪個attribute對結果影響最大？

我認為capital gain是影響最大的attribute，經由測試每次只用單一個特徵訓練與預測，發現只用Capital Gain 即可達到 0.81 以上的準確度(尚可接受)，至於其他特徵如果單一使用，則準確度都只有大約 0.74 ~ 0.77(幾乎跟用猜的一樣)。