Homework 2 Report - Income Prediction

學號：b04505021 系級：工海三 姓名：黃廉弼

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

	ACC on public score	ACC on private score
Generative model	0.76474	0.76280
Logistic Regression	0.84815	0.84571

明顯可看出 Logistic Regression 的 Performace 較 Generative Model 好上許多，也許是跟老師上課提及的，generative model 比較適合在 data 量比較少的時機；Logistic Regression 比較適合在 data 量比較多的時候，因此在這邊 Logistic regression 會有較佳的 ACC。

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

主要以上課教到的 Logistic Regression 實作，搭配 normalization、regularization、batch normalization 等技巧。

而 Loss 以 cross entropy 計算，Activation function 則使用 sigmoid function，並未使用任何 optimizer(adagrad、adam...)，即僅用常數的 learning rate。最終 ACC 落在 85.5%上下。

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)

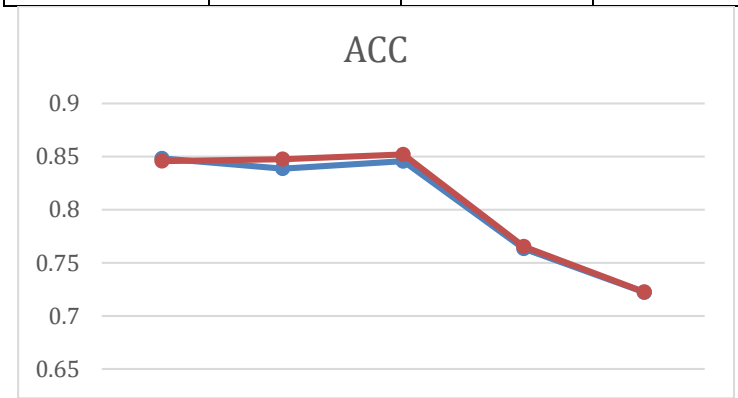
	ACC without normalization (public/private)	ACC with normalization (public/private)
Generative model	0.83736/0.84250	0.76474/0.76280
Logistic Regression	0.79668/0.78638	0.84815/0.84571

可以看出來，在同樣的 Learning Rate 以及 epoch number 下，Logistic Regression 中，feature normalization 明顯對 perfoemance 有極大的幫助。然而 Generative Model 卻呈現相反結果，我猜測因為 generative model 是

以 Guassian Distribution 做運算有關，如果先做 normalization 影響整體分布反而會使其預測的結果較差。

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)

lambda	0	1e-5	1e-3	1	10
ACC(private)	0.84815	0.83859	0.84559	0.76329	0.72214
ACC(public)	0.84571	0.84754	0.85196	0.76523	0.72248



其實不難看出，regularization 對於整體 performance 並無太大幫助，甚至當 lambda 大於 1 後有拖累 performance 的狀況。

雖然說，經過 shuffle，performace 有所浮動在所難免，但整體趨勢看來 regularization 確實在此處是軍無用武之地。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

根據 Train 出來的 weight 經排序後得到前五大 weight 的資料及其 weight 分別為:

attribute	|weight|(以科學記號表示)
capital_gain	1.45402344e+00
education_Preschool	1.11164049e+00
education_num_1	1.11164049e+00
marital_status_Married-civ-spouse	7.19313752e-01
marital_status_Never-married	5.28197746e-01

由以上結果，我認為 capital_gain 是對結果影響最深的 attribute。