Homework 2 Report - Income Prediction

學號：b05901033 系級：電機二 姓名：莊永松

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

logistic regression較佳。雖然在只餵raw data未做normalize時，generative model仍然可以保有0.84左右的準確率，勝過logistic regression(0.72)許多。但在經過normalize，並且實作adam，加上regularization term後，logistic regression最後的表現較好。
推測是因為generative model是close form，無法針對一些特別狀況做微調，而logistic regression對於raw data雖然表現不好，但在data經過處理並且使用較佳的gradient descent方法後，能走到loss function低點的地方，對資料能夠更加貼合。
以下是kaggle成績比較。(data有將continuous的term增加高次項(2次項~6次項)以及ln(1+x)項)

	public score	private score
generative model	0.85933	0.85407
logistic regression	0.86130	0.85775

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

使用scikit-learning內建的logistic regression
regularization 使用 L1 penalty
data有將continuous的term增加高次項(2次項~6次項)以及ln(1+x)項
kaggle public score: 0.86130
kaggle private score: 0.85959

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)
(此題的score是切validation set五次平均的結果，非kaggle成績)
(1)使用standardlize (減去mu，除以sigma)
generative model score: 0.84120
logistic regression score: 0.85116 -> 表現最好
(2)使用max-min normalize (x-min)/(max-min)
generative model score:0.84126 ->跟上面差不多，因generative model較不受normalize影響
logistic regression score: 0.84532 ->下降不少，推測是max-min normalize易受極端值(outlier)影響，故標準化效果受影響，梯度下降無法達到最佳
(3)不做normalize
generative model score: 0.84126 -> 完全不受影響。
logistic regression score: 0.72584 -> 偏爛，推測沒做normalize，很難用gradient descent走到最下面，且過程中loss一直跳動不定，看得出原本的learning rate對此data而言過大。

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)

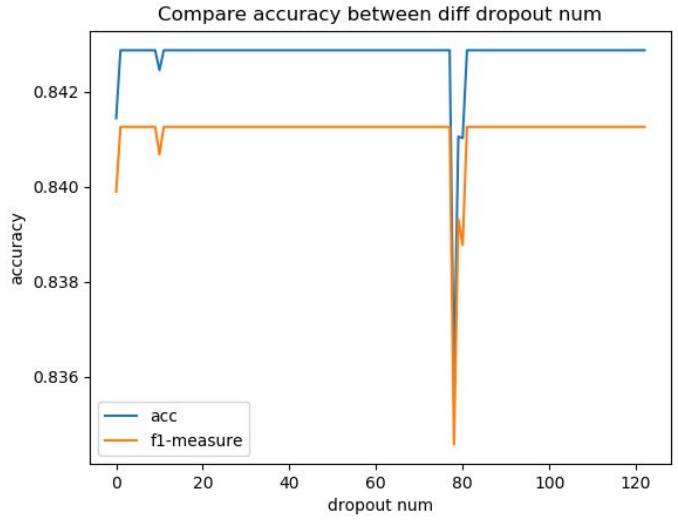
(lambda = 0.001)	**public score**	**private score**
L2-regularize	0.86031	0.85763
L1-regularize	0.85995	0.85751
不做regularize	0.85577	0.85751

在我實作的logistic regression中，regularization不論是L1、L2都有不錯的效果，在我增加高次項到六次項時，能夠避免overfitting，使得增加高次項對預測結果有所幫助。
使用L1-regularization時，會讓目標weight變得較為sparse，也就是有最多的0，能夠用最少量的vector就組合出正確答案，而在這次作業data之中，許多feature是useless的，作L1正好能夠把他們去掉，因此效果不錯。在我使用sklearn的logistic regression時，L1的效果(~0.86)遠大於L2(~0.72)，但在我自己的model上沒觀察到這件事。推測是sklearn可能還有使用其他最佳化方法，導致使用L2時易受到outlier影響而效果變差，而我的model較為簡單，L2的效果不受影響，反而略比L1的成績好一小點。

5. (1%) 請討論你認為哪個attribute對結果影響最大？
我做了dropout的實驗(使用generative model)，輪流將某一個attribute拿掉再用generative model來測試，發現大部分的項對結果幾乎無影響。有明顯影響的只有5項，影響大小排列依序為：

```
[78]'capital_gain'   => 0.83588
[80]'hours_per_week' => 0.84102
[79]'capital_loss'   => 0.84105
[0]'age'             => 0.84144
[10]'fnlwgt'         => 0.84245
```

其他attribute拿掉後，結果都是 `0.84287` ，影響不大
因此我認為**capital gain對結果的影響最大**，也十分合理，因為投資收入跟年收入有很大關係。



※對這次作業的額外建議：我認為用F1-measure來做kaggle上的評分依據會比較準確，畢竟這次資料的y_label中，0占了多數，用準確率來評分較無法真正反映model好壞。