Homework 2 Report - Income Prediction

學號：b03901098 系級：電機四 姓名：王建翔

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

（兩種做法皆有經過normalize pre-processing）

	Average acc
Generative model	0.844045
Logistic regression（with best_feature）	0.85836

從上表可以得知Logistic regression可以得到比較好的準確率，但是generative model的收斂速度較快，一但樣本數較多，可以較快的方式收斂到true model。

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

實作出的best model是使用logistic regression，並對原本123-dim的資料做處理，第一步是先找出個別attribute對準確率判斷的影響，後來發現index 3,4,5，Age，Capital gain，Capital loss，fnlwgt為影響最大的幾項，所以將其組成"feature"，並對其一次方、二次方三次方和原本的123-dim做concatenate成為148-dim的feature去做training；Weight則是用adagrad做更新。

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logistic regression為training方式

	Average acc
Without normalization	0.85744
After normalization	0.86061

可以發現在經過feature normalization之後，準確率有明顯的提高；因為這次全部的feature有很多因為是one-hot coding所以是0或是1的情況，但是這次我選用的attribute，如Age, fnlwgt等等都是數字比較大的，經過normalization會使得feature彼此的影響必較平均。

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	Public	Private
100	0.85712	0.85554
10	0.85687	0.85640
1	0.85702	0.85761
0.1	0.85773	0.85702

Regularization能夠讓我們能夠有更smoother的function來fit並最小化誤差，發現有做正規化的影響並不大，可能在一開始train出來的model就已經算是fitting，而正規化之後有些overfitting的情況。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

使用logistic regression分析

將每一個feature都先各自獨立取出，並和原本的全部123維合併之後做training，並且對每一個feature的準確率做比較，而發現影響最大的分別是：Age,Capital_gain, Captital_loss,fnlwgt，而其中Age的影響又是其中最大的。