Homework 2 Report - Income Prediction

學號：r06921077 系級：電機碩一 姓名：黃詩凱

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

   很明顯地，logistic regression準確率較佳，generative model的分數是0.84557；logistic model的分數則是0.85810。

   老師上課時也有說，一般來說generative model不一定有比較好的結果，因為他是假設在某個機率模型下的狀況(腦補)，在資料量少的時候是有可能會比discriminative model還要好的。

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

   使用手刻logistic regression，有用Adagrad和Normalization，在kaggle上的分數為0.85810，訓練參數: learning rate=10；iteration=30000，其他部分幾乎都跟作業一差不多，然後bias合併到w中一起算。

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)

   特徵化之後明顯結果較佳，原本過不了strong baseline，改用normalization就上去了。

   主要可能是因為feature之間差異過大，有的值很大，有的值很小，需要解決這個問題。

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)

   regularization對我的model影響不大，我嘗試過調整regularization前的rate(rate=0.1, 1, 2, 10, 20, 100都試過)，只有在rate=2時，分數有比較高一些，其他狀況下反而排名往下降，但也有可能只是overfitting public testing data就是了。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

   我的測試方法：
   先測出全部feature一起下去train的結果
   再來如果要測feature i，就把feature i單獨去掉，看分數的差異來判斷影響力。

使用全部feature: 0.85810
去掉feature 0: 0.85565
去掉feature 1: 0.85700
去掉feature 2: 0.85773
去掉feature 3: 0.85773
去掉feature 10: 0.85405
…
去掉71~75(race): 0.85687
去掉feature 76,77(sex): 0.85786
去掉feature 80(hours_per_week): 0.85429

雖然沒有全部試，但試了好幾個feature，最後發現feature 10去掉之後，分數降最多
而hours_per_week也很理所當然地是降第二多的
所以我覺得影響最大的feature試fnlwgt