Homework 2 Report - Income Prediction

學號：b05611033 系級：生機二 姓名：杜杰翰

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

   Logistic regression 的準確度較佳。在沒有增維的情況下，實作出來有做 nomalization 的 linear regression 的準確度有 0.85，沒做 nomalization 的有 0.8，但是 generative model 跑出來結果只有 0.79。我想是因為 generative model 依賴樣本比例，若 training data 中大於 50k 的比例較多，則 generative model 的結果就會比較偏向大於 50k，反之亦然。而 logistic regression 不注重樣本比例，反而較注重各項 feature 對結果的影響，所以在這種非黑即白的分辨之中比較不會因為 data 的數量關係而影響結果，因此會較 nomalization 準確。

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

   我使用手刻的 logistic regression，並使用 adagrad 將全部數據跑 60 次。
   一開始我先將數據標準化，然後針對原始數據中不是非零即一的 feature 進行 181 次方及 cos、sin、tan、arctan 的增維，之後再進行一次標準化。我還將資料分 1/5 出來做 validaiton，並使用其他 4/5 做 training，然後才開始進行 logistic regression。

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)

   我在沒有增維的情況下比較有無 normalization，在有做的情況下正確率可以到達 0.85，而沒做的話只有 0.8。我想這是因為這次 data 中大多是的 feature 都是 0 與 1，若沒有做標準化這樣 0 跟 1 的影響過於巨大，但這不應該只是有與沒有而已，若剛好 0 與 1 的人是一半一半，這樣人群會被極端的分為兩部分。而若有做 normalization 的話，則會視 0 與 1 的比例，進行數字的調整，使得這個 0 與 1 不再只是有與沒有的差別，並以更精確的數字顯示了他們之間差異，從而可以讓 training 出來的結果更為準確。

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)

   我在有做 normalization 的情況下，進行了有 regularization 與沒有的比較，發現在 hw1 可以有效增進成果的 regularization，在這個 case 中反而會造成反效果。在沒做的情況下我的正確率有 0.85，然而做了以後我的正確率卻只剩下 0.8。而在我拿了 best 去做 regularization 之後，我的 best 的結果從 0.859 變成了 0.77。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

在各項 feature 中，native_country_United-States 的 weight 最大，因此我認為「是否為美國人」對收入是否大於 50k 的影響最大。