學號：B06507007 系級：材料二 姓名：王致雄

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

   Generative:準確率約 84.4%　Logistic: 準確率約 85.6%

   符合課堂上 Logistic 優於 Generative 的評論。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

   我有試著使用 pytorch 用 DNN 預測，但是結果均無法超越手刻的 Logistic regression，我也很好奇為什麼。所以我的 best model 就是 Logistic regression。其中把性別二元化，並把 fgnwt 給刪除(看他的說明我覺得沒什麼用)。然後將所有為整數的 feature 都除以整筆資料的最大值(效果相當於標準化)。

3. 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

   標準化能讓 feature 大小相近，使 weight 的大小能反映 feature 的重要性，且收斂較快，加速 trining。並且若 feature 過大，training 的時候很容易報錯，標準化可以有效防止這點。

   助教的 generative sample code 是將所有 feature 包含 one hot encoding 全部標準化，而我嘗試了只對整數的部分作標準化，可以把準確率再提高一點點

   實際上若不使用我上述的方法改使用標準化，實作了幾個，準確度約落在 85.3~85.4%附近，與上述方法相差不大。

4. 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。

   嘗試了 Lambda=0.1~0.0001 我認為看不出什麼效用，且最終準確率也相差不大(在 0.1%以內)。我認為原因可能是這次的 regression iteration 較少(約 100)，比較不會有 overfit 的問題。

5. 請討論你認為哪個 attribute 對結果影響最大？

   看了一下 Logistic regression 的 weight，大多在-5~5，唯有 capital gain 這項高達 27，可見其為非常重要的 feature。想想也相當合理，若一個人擁有很多資產，則他的年收入就很有可能較高。