學號：B05902128 系級：資工三 姓名：鄭百凱

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

   在沒有作任何處理，單純將全部的 data 丟進去 train，generative model 的結果會比 logistic regression 好，generative model 在 public 的結果可以到 0.846，可以過 public simple，而 logistic regression 則不行。但是在 normalize 之後，logistic regression 的結果會比 generative model 要來的好。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

   資料處理的部分，我將國籍除了美國以外的欄位都刪除，人種也只留下白人，並將學歷都刪除，取而代之的是原始資料中的 education_num 這個欄位，還有把一些可以被其他欄位線性組合出來的欄位拿掉。然後將不是 0/1 的欄位取 1 到 40 次方之後 normalize 用 sklearn 下去 train。在 public 拿到 0.85859，private 拿到 0.86021。

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

   同樣在 learning rate 為 1、iterate 10000 次、起始值為 0 的狀況下，沒有 normalize 的準確率只有 63%, normalize 則有 84%。

   沒有 normalize 的時候，結果容易受到一些變動太大的參數影響，在 training 的過程中，error rate 也是忽大忽小。此外，只要稍微改變起始值和 iterate 次數，對結果可能就有很大影響，在其他參數的情況下，我做出來沒有 normalize 最好的結果可以到 83%。

   有 normalize 的話，training 的過程中 error 會穩定的逐漸變小，出來的結果比較有保障。也因為數字較小，計算速度比較快，training 所需的時間較短。

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。

   我測試了 lambda = 0.01、0.1、1、10、100，其中表現最好的是 lambda = 0.1，但這五個的差距並不大，大概是正負 0.2%左右的差距。我認為在這個 task 上，regularization 並沒有太大的幫助。

5. 請討論你認為哪個 attribute 對結果影響最大？

   我將每個 feature 對應的 weight 印出來，其中對薪水的影響較大的是 education_num(在原始資料中)、sex、age、capital_gain。然後分別將他們整欄拔掉再

丟到 kaggle 去看 accuracy，發現 accuracy 下降最多的是 capital_gain，其次是 education_num。age 和 sex 雖然也有一定的影響，但相較於前兩者，影響力要小的多。

因此我認為最重要的 attribute 是 capital_gain，其次是學歷。這兩個 attribute 應該也是直觀上在這麼多 feature 中最能影響薪水的 attribute。