學號：R06922048 系級： 資工碩一 姓名：陳柏堯

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

答：

在兩者 model 上使用相同的 features 來比較：

* 使用 Generative Model，將全部 X_train 的 features 放入之後得到的結果
  public score:0.84471，private score:0.84203，valid score: 0.83550，
  Train score: 0.84177
* 使用 Logistic Model，將全部 X_train 的 features 放入之後
  (因為第 0、3、4、5 的數值較大，若直接做 sigmoid 會 overflow，所以針對這 4 個 feature 做 normalization)
  public score:0.85393，private score:0.85100，valid score: 0.854846，
  Train score: 0.8533321

因此比較發現 Logistic regression 在相同情況下表現較佳。

Note： Validation 和 Training data 是從原始 data 中 3：7 random shuffle 之後切出來的。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

以我選擇的兩個 Kaggle model 中 private score 較高者 作為 best model，使用的是 Logistic Regression，將所有 feature 放入 train，並且對第 0、1、3、4、5 的 feature 做 normalization（分別減掉其平均數再除以標準差），learning rate 為 0.00005，iteration (epoch) 的次數為 15001，初始值的 weight 的 random 範圍在-1e-5 到 1e-5 之間，得到的結果是 public score:0.85393，private score:0.85100。

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。

答：

* 在 Generative Model 中，對第 0、1、3、4、5 feature 做 normalization 的話，得到的 Train Accuracy 為 0.763732，Validation Accuracy 為 0.759190；而完全不做 normalization 的話得到的 Train Accuracy 為 0.84177，Validation Accuracy 為 0.83550。由此發現，對於 Generative Model 不做 normalization 比較好，我推測因為這些數值大的 feature 在 Generative Model 中具有將不同 data 數值拉開的特性，若 Normalization 之後則會減弱這種特性，導致結果不好。
* 在 Logistic Model 中，因為大部分數值都是 0 或 1，因此這些我認為不需要做 normalization。若對於數值不是 0 或 1 的第 0、1、3、4、5 feature，做

normalization，Training 如第一題一樣 feature 全部放入，得到的 valid score: 0.854846, Train score: 0.8533321，若不對這些 feature 做 Normalization 的話，在 trianing 過程中就會 overflow，我想這是由於在 sigmoid function 中值太大而導致出現 nan 的問題。

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。

答：

按照 hwl 當中的方法對 loss function 及 gradient 加上 regularization 項

lambda = 0， private=0.85100, public=0.85393, validation: 0.83550, train: 0.84177

lambda = 0.01， private=0.85124, public=0.85393, validation: 0.85228, train: 0.85411

lambda = 0.1， private=0.85100, public=0.85393, validation: 0.85597, train: 0.85196

經過比較之後發現，使用 regularization 對於準確率影響不大，但是仍然會有幫助，準確率有小小地提高。

5.請討論你認為哪個 attribute 對結果影響最大？

* 在實作中的觀察，以 Generative Model 來做 feature 的調整，一開始全部 feature 放入 train 得到的結果是 public score:0.84471, private score:0.84203, valid score: 0.83550, Train score: 0.84177，後來經過各種調試，再將第 0、3、4、5 的 features 開 0.5 次方、1.3 次方，放入到 model 中，得到 public score: 0.85835, private score: 0.85001，valid score: 0.84972, Train score: 0.85701，準確率得到進步，表示這些 features 對於結果有很好的幫助。

* 利用 Generative Model 對所有 features 作 training 得到一個結果，然後再一個一個分別拿掉每一個 feature，去觀察哪一個 feature 去掉之後影響的結果最多，得到的結果是第 2、4、1、0、5、3、91、81、97 這些 features，其對應的項目是分別是 sex, capital_loss, fnlwgt, age, hours_per_week, capital_gain 等。