學號：R06922113 系級： 資工碩一 姓名：陳宣伯

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

答：

以下計算 Accuracy 的方式為將 Public 與 Private 的分數做平均，得到結果如下表：

Model	Accuracy
Generative model	84.405%
Logistic regression model	84.540%

兩個模型的輸入都有經過 normalize，實作方法可參照第 3 題，而我的 Logistic model 參數設定為以下：

- 取了每個特徵值的 1 次方 + 2 次方
- Regularization: lamda = 0
- Gradient descend algo. : Adam

可以看得出來 Logistic regression model 的 Accuracy 高出 Generative model 一些些，但執行上 Logistic model 必須 train 將近 5 分鐘，generative 則是不用 train 幾乎不到半分鐘就把答案算出，因此若同時考慮 training 的時間以及 Accuracy，generative model 的表現較 logistic model 來的好

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我的 best model 是用 Keras 的 ann 訓練的，其參數如下：

- 訓練樣本數：32561 個
- 特徵：106 個一次項 + 106 個二次項
- 正規化：$x' = \frac{x - \bar{x}}{\sigma}$， $\bar{x}: mean$; $\sigma: standard\ deviation$
- NN 架構
  - Input layer : 212 nodes
  - Hidden layer1 : 60 nodes (Activation : relu)
  - Hidden layer2 : 30 nodes (Activation : relu)
  - Output layer : 1 nodes (Activation : sigmoid)
  - Dropout rate : 20% on the input layer and hidden layer
  - Gradient descend algo. : Adam

以上述的參數訓練的結果，public 與 private 的平均辨識率為 85.412%，比第 1 題討論的 Generative model 或 logistic model 都來得佳，但是 public 的分數與 private 的分數相差了 0.846%，相較於在 logistic model 上的實驗來的大許多，所以有一些 overfitting 的現象，之後實驗可能要把 dropout rate 調高

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：

以下計算 Accuracy 的方式為將 Public 與 Private 的分數做平均，得到結果如下表：

Feature normalization	Accuracy
True	84.540%
False	79.6505%

可以看到有沒有將輸入做 feature normalization 是非常關鍵的，兩者 Accuracy 相差了將近 5%，所以其他所有的實驗我都會先將我的輸入做 normalization，原因是有非常多的 feature 是 0，對於這樣的值 learning rate 必須要調的極小，但 gradient descent 就很容易卡住無法更新或更新速度極慢，且結果也不甚理想。

我這邊對輸入做 normalize 的方法就是以下式子：

$$x' = \frac{x - \bar{x}}{\sigma}, \quad \bar{x}: mean \ ; \ \sigma: standard\ deviation$$

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：

Lamda	Accuracy
0.1	84.540%
0	84.540%
1	84.422%
10	84.03%

非常合理隨著 Lamda 的增大，Accuracy 會下降，但是將 lamda 增加 100 倍，Accuracy 也不至於損失太多，故本次實驗 regularization 的參數設定在 10 以內不會影響結果太多

5.請討論你認為哪個 attribute 對結果影響最大？

這裡我只討論 logistic model 的複雜度、標準化、正規化來討論，複雜度的部分當我取所有特徵的 1 次方以及 2 次方去計算，得到的結果如下表：

Feature complexity	Accuracy
1 次方	84.422%
2 次方	84.540%

可以發現在 2 次方以下的模型複雜度對模型結果並無顯著的影響。

再來是標準化(feature normalization)以及正規化(regularization)，這兩個部分的討論已在報告中的第 3 題、第 4 題做過討論

總結來說，我認為有沒有將輸入做特徵化影響模型結果非常大，可以看到只要有做 feature normalization，Accuracy 都是 simpleBaseline 起跳