學號：B03901165 系級： 電機四 姓名：謝世暐

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
答：
在 Validation set 上，logistic 得到的正確率是 0.85278，而使用 Gaussian、shared covariance 的 generative model 的 0.84369。也就是 logistic 的準確率較佳，在 kaggle 上 public 準確率略有不同但也是 logistic 較佳，和課堂上的結論也相同。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
實作中結果最好的是使用 Xgboost 這個 package，以 Gradient Boosted Tree 作為 model，max_depth 設為 10，以 3 個 fold 的 cross validation，在最低的 error 停止訓練 (early stopping)。最後得到在 Validation 的準確率是 0.86801，在 kaggle 上 public 的準確率則是 0.86916，好過其他試過的 model。(另外試過用 Random Forest 準確率在 kaggle 為 0.86523、keras 建立的 NN 準確率為 0.86191)

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：
Logistic: 沒有作 normalization 的話，y 很容易因為 overflow 而變成 1，導致 1-y 取 log 之後變成-inf。有作 normalization 的話，使用 2000 個迴圈，step = 0.1，準確率是 0.85278。但沒作 normalization，必須用小一點的 step(0.0001)，使用 20000 個迴圈，也只能得到 0.81 的準確率，表示 normalization 對 training 的影響相當大。
Generative: 由於 Generative model 沒有 overflow 的問題，所以在實驗中無論有沒有做 normalization，都得到了 0.84369 的準確率，表示 normalization 對準確率基本沒有影響。

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：
沒有使用 regularization 的準確率如前所述是 0.85278，我實作了 L1 和 L2 的 regularization，準確率結果如下:

Lambda	1	10	100
L1 regularization	0.85284	0.85315	0.85278
L2 regularization	0.85271	0.85241	0.85174

上面的數字皆為 Validation 的準確率，L2 在 lambda 增加時準確率卻逐漸下降，L1 則是先升後降。從結果來看，L1 regularize 且 lambda 為 10 的結果最好。

5.請討論你認為哪個 attribute 對結果影響最大？

答：

以下是將 Logistic Regression (with regularization)的其中一筆特徵拿掉後的結果:

Age	Fnlwgt	Sex	Capital gain	Capital loss
0.85186	0.85272	0.85309	0.83798	0.85008
Hours per week	Workclass	Education	Marital status	Occupation
0.85081	0.85088	0.84203	0.85217	0.84682
Relationship	Race	NativeCountry		
0.85309	0.85309	0.85296		

就以上的結果，會發現使用全部特徵(0.85315)還是比拿掉任何一個特徵來的好。如果將影響最大的特徵，視為拿掉後會讓準確率下降最多的特徵，則 Capital gain 會是對結果影響最大的，其次是 Educationi，再其次是 Occupation。