學號：R05921087 系級： 電機碩一 姓名：鍾智堅

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：
首先我刪除了有關 marital_status 的 attribute，并對 X_train 和 X_test 做 normalize 處理，然後找出兩個 sigma 和共同 sigma 的矩陣以找出模型。其準確率比起 discriminative model 差一些些，因為其訓練方式所規劃的數據分佈邊界為模糊，因此在這次的資料庫中表現比較差。
其原有的程式準確率為大約 0.842265

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：
這次的作業主要是用此模型去找最佳化的解答。我有刪除了有關 marital_status 和 education-num 的 attribute。其訓練方式也是把數據經過 feature normalization 后，再用 batch gradient descent 去訓練。我也在模型上加了 regularization 防止 overfit 的問題。
其準確率大約為 0.858283

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：

```
# feature normalization with all X
X_all = np.concatenate((X_train, X_test))
mu = np.mean(X_all, axis=0)
sigma = np.std(X_all, axis=0)

# only apply normalization on continuos attribute
norm_index = [0, 1, 3, 4, 5]
mean_vec = np.zeros(X_all.shape[1])
std_vec = np.ones(X_all.shape[1])
mean_vec[index] = mu[index]
std_vec[index] = sigma[index]

X_all_normed = (X_all - mean_vec) / std_vec
```

在 discriminative model 中，未加 normalization 的準確率為 0.84073，而加了 normalization 后的準確率為 0.858283

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 = 0.0000001
w_grad = (1.0/m) * np.sum(-1 * X * (Y - y).reshape((batch_size,1)), axis=0) + (lamda/m)*(w.T.dot(w))
w2_grad = (1.0/m) * np.sum(-1 * X**2 * (Y - y).reshape((batch_size,1)), axis=0) + (lamda/m)*(w2.T.dot(w2))
b_grad = np.sum(-1 * (Y - y))
```

未加入 regularization 前其準確率為 0.858037，加了后其準確率為 0.858283

5.請討論你認為哪個 attribute 對結果影響最大？

經過測試后，發現有兩個 attribute——Capital-gain 和 Relationship 對結果影響最大。

在未刪除前任何 attribute，其準確率為 0.858283，而分別刪除后其結果展示在以下圖中。

刪除 Capital-gain 后的準確率:

```
linux@linux-T460s:~/Documents/MachineLearning/HW2$ python2.7 hw2_logistic.py X_t
rain.csv Y_train.csv X_test.csv ./result/prediction.csv
valid accuracy in epoch50: 0.840934
valid accuracy in epoch100: 0.843697
valid accuracy in epoch150: 0.843237
valid accuracy in epoch200: 0.843390
valid accuracy in epoch250: 0.844004
valid accuracy in epoch300: 0.844158
valid accuracy in epoch350: 0.844465
valid accuracy in epoch400: 0.844465
--- 45.3366379738 seconds ---
```

刪除 Relationship 后的準確率:

```
linux@linux-T460s:~/Documents/MachineLearning/HW2$ python2.7 hw2_logistic.py X_t
rain.csv Y_train.csv X_test.csv ./result/prediction.csv
valid accuracy in epoch50: 0.835406
valid accuracy in epoch100: 0.836174
valid accuracy in epoch150: 0.836327
valid accuracy in epoch200: 0.836174
valid accuracy in epoch250: 0.836174
valid accuracy in epoch300: 0.836174
valid accuracy in epoch350: 0.836481
valid accuracy in epoch400: 0.835560
--- 47.2230510712 seconds ---
```