學號：B02902044 系級：資工 姓名：陳映紅

1.請說明你實作的 generative model，其訓練方式和準確率為何？

- 特徵：使用所有 106 維的一次方的資料，沒有實作輸入特徵標準化。
- 訓練方式：以 Y_train 值將所有 X_train 分為兩個類別，Y_train = 1 時為第一類，Y_train = 0 時為第二類。利用分類好的訓練資料和下列方程式算出 w 及 b 後，再將 z = w • X_test + b 代入 sigmoid 方程式，求得每筆測試資料屬於第一類的機率，大於 0.5 則輸出 1，反之則 0。

$$z = \underbrace{(\mu^1 - \mu^2)^T \Sigma^{-1}}_{\boldsymbol{w^T}} x - \underbrace{\frac{1}{2}(\mu^1)^T \Sigma^{-1} \mu^1 + \frac{1}{2}(\mu^2)^T \Sigma^{-1} \mu^2 + ln\frac{N_1}{N_2}}_{\mathrm{b}}$$

- 準確率：Kaggle Public Score 為 0.84165，Private Score 為 0.84658

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？

- 特徵：使用所有 106 維的一次方及二次方資料，有實作輸入特徵標準化。
- 訓練方式：先計算所有訓練和測試資料的一次方及二次方值，再一併做特徵標準化，使訓練和測試資料的標準化統一。利用 Gradient Descent 找到 w 和 b 後，將 z = w • X_test + b 代入 sigmoid 方程式，求得每筆測試資料屬於第一類的機率，大於 0.5 則輸出 1，反之則 0。
- 準確率：Kaggle Public Score 為 0.85627，Private Score 為 0.85788

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。

我在計算 sigmoid 方程式時使用到 math.exp()函式，函式 input 限制為[-709, 709]，超過此限即溢位。第 2 題的 model 在沒有實作特徵標準化時，程式會產生 overflow error，因此我另外使用 numpy.clip()，將過大或過小的 input 調整為 709 或-709。下表為第 1 題及第 2 題的 model 加上特徵標準化後的結果：

	Kaggle Public Score	Kaggle Private Score
Generative Model 有標準化	0.84165	0.84621
Generative Model 無標準化	0.84165	0.84658
Discriminative Model 有標準化	0.85627	0.85788
Discriminative Model 無標準化	0.79656	0.79671

由上表可知，Generative Model 加上標準化後準確率並無明顯差異，但 Discriminative Model 加上標準化後因溢位處理所以使準確率顯著降低。

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。

我以第 2 題的 model 另外加上正規化，結果如下表：

Lambda	Kaggle Public Score	Kaggle Private Score
10	0.85614	0.85800
100	0.85418	0.85469
1000	0.84582	0.84633
10000	0.83120	0.83602

由上表可知，lambda = 10 對 Discriminative Model 有些微幫助，但往後 lambda 值越大，model 整體表現越差。

5.請討論你認為哪個 attribute 對結果影響最大？

我將各種不同 feature 去掉後以第 2 題的方式訓練模型，得到下表結果：

去掉哪種 feature	Kaggle Public Score	Kaggle Private Score
age	0.85295	0.85321
fnlwgt	0.85528	0.85690
sex	0.85676	0.85788
capital_gain	0.84324	0.84265
capital_loss	0.85369	0.85456
hours_per_week	0.85319	0.85800
workclass	0.85565	0.85579
education	0.85012	0.85198
marital-status	0.85627	0.85825
occupation	0.85012	0.85284
relationship	0.85663	0.85849
race	0.85688	0.85776
native-country	0.85565	0.85899

由上表可知，刪去 feature 後，準確率會有些微上升或下降。而將 capital_gain 抽掉後的準確率改變（下降）最多，因此我認為這是影響最大的 attribute。