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。不過，此方法的準確率很差，Kaggle Public Score 只有 0.79656，而有作標準化、沒有使用 clip() 時的 Kaggle Public Score 為 0.85627，由此可見標準化之重要性。

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，當 lamda = 100 時 Kaggle Public Score 為 0.85418，當 lamda = 1000 時 Kaggle Public Score 為 0.83120，可見 lamda 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？

我觀察第 2 題的 model 的 w，發現絕對值最大的為 w[0] = 3.19617236，第二大的為 w[106] = -2.6915993，前者為「年齡」的一次項，後者為「年齡」的二次項，可見在這個 data set 中，年齡對於結果的影響最大。