學號：R04945008 系級：生醫電資碩二 姓名：黃思翰

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

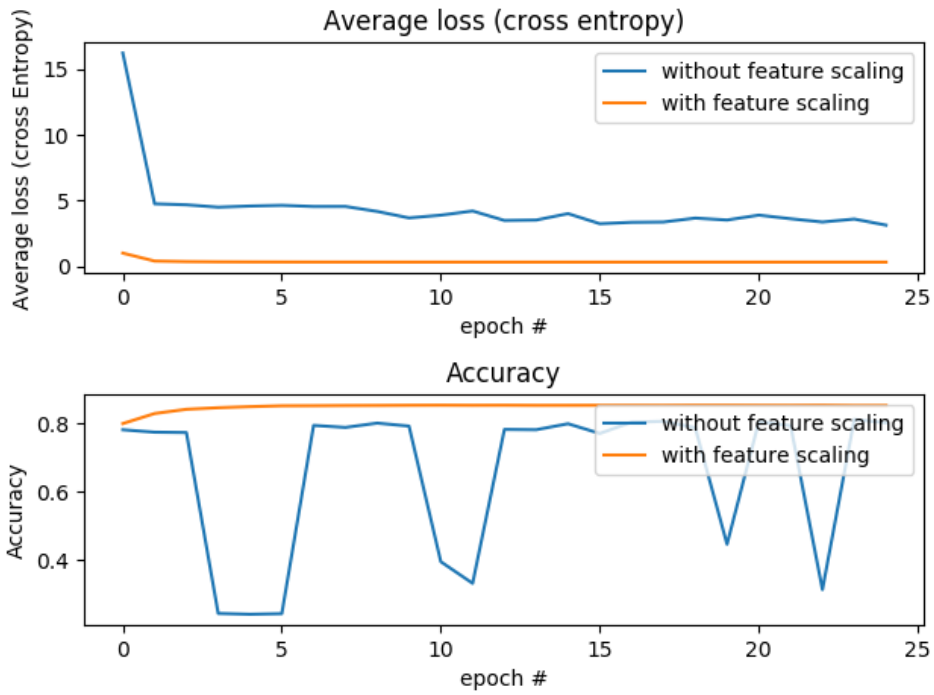
答：使用助教提供的 106 維訓練資料，並對連續型態的資料做 feature scaling 後，我假設此 data 為高斯分布，並計算出年收入大於 50k 和小於 50k 兩類資料的 mean 及 covariance matrix 後依比例算出最終的 covariance matrix。接著在預測時先計算出 covariance matrix 的反矩陣，並利用老師上課推導的高斯分布的預測模型公式，計算出對應的 w 及 b，最終將測試資料乘上 w 並加上 b 後通過一個 sigmoid function，來得到預測結果，而我在 kaggle 上的 public 準確率為 0.841

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

答：實作 logistic regression，使用助教提供的 106 維訓練資料，進行訓練前先將連續型態的資料做 feature scaling，以助教提供 106 維資料為例即 age, fnlwgt, capital_gain, capital_loss, hours_per_week 此五維的特徵以提高模型準確率。在 kaggle 上突破 public strong baseling 的訓練條件如下：每 50 筆為一 batch，總共訓練 25 個 epoch，每個 epoch 開始前會將訓練資料的順序隨機打亂，learning rate 為 0.002，optimizer 則是實作 adam 來提高訓練的速度，透過 gradient descent 的方式來降低 average loss (cross entropy)，經由反覆訓練後，計算出 w 和 b，最終將測試資料乘上 w 並加上 b 後通過一個 sigmoid function，來得到預測結果，最後在 kaggle 上的 public 準確率為 0.855。

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

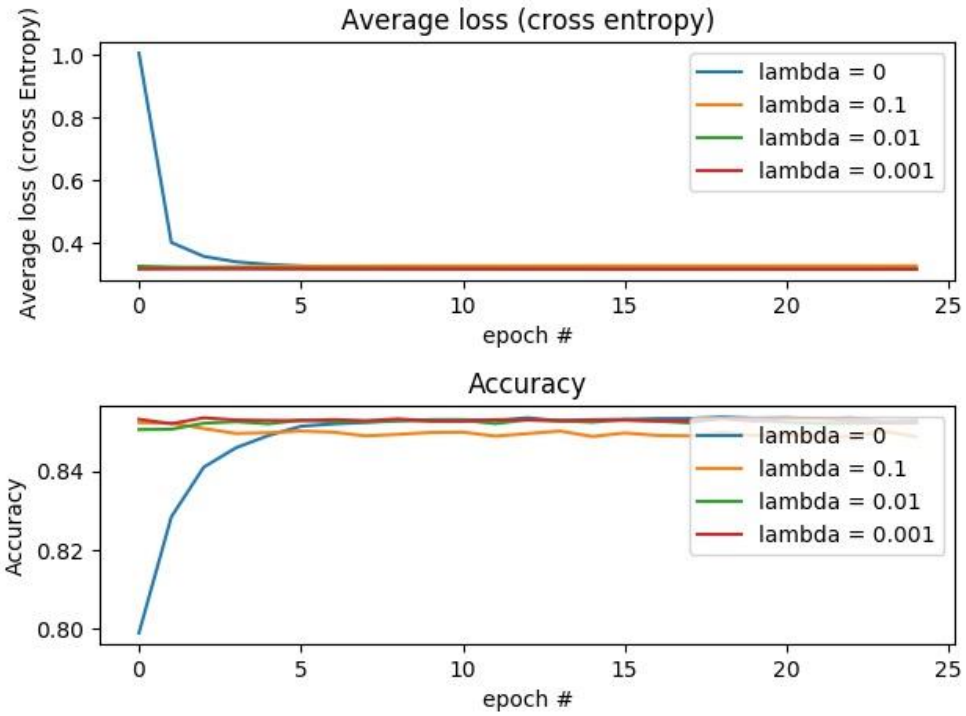
答：本題使用助教提供的 106 維訓練資料，並將連續型態的資料減去平均值再除以標準差做 feature scaling，並比較了有無 feature scaling 對於 average loss 和準確率的影響，從下圖可以觀察到，加入 feature scaling 對於模型的準確率有顯著的提升。



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：本題我實作了 L2 正規化，於 loss function 中多加了 $\lambda * \sum_{i=1}^{N} \omega_i^2$ 項，並比較了不同 lambda 對於 average loss 與 accuracy 的影響，以下數據均為使用助教提供的 X_train、

Y_train 的結果。由下圖中可觀察到當 lambda = 0.01 和 lambda = 0.001 時於 training data 所得到的準確率已經提升到上限，所以我就沒繼續縮小 lambda 了。



5.請討論你認為哪個 attribute 對結果影響最大？

答：Logistic regression 可以寫成數學式 sigmoid($\sum_i w_i x_i + b$)，因此此題詢問哪項 attribute 對結果影響最大我認為可以從 sigmoid function 的輸入來探討，越重要的特徵所算出來的$w_i x_i$其值應該越大，而其正負則是影響分類，因此我使用助教提供的 106 維特徵並做完 feature scaling 後，先計算出 32561 筆的資料在 106 維中各個特徵的標準差，並將此結果乘上我所訓練出來的模型的 w，以此來觀察哪項 attribute 對結果影響最大。從下圖結果觀察到 capital gain 此項 attribute 在這樣的計算下有最大的影響力。