學號：R05921064 系級： 電機所碩一 姓名：羅嵩詠

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：將 X_train 內，所有資料依年收入是否大於 50k 分類。再做特徵標準化。之後個別找出它們的 mean 以及 covariance。把兩個 covariance 矩陣加權平均，計算出一個新的矩陣。再代入教授上課投影片上的公式，得到 w 與 b。
準確率：0.76032 (kaggle)

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：X_train 內原有 106 組特徵，拿掉 fnlwgt，並將四組帶有連續數值的特徵，經過等級劃分，轉換成 01(是/否)離散型式。最後共有 115 組輸入特徵。age (青年 (0-25), 狀年 (26-45), 中年 (46-65), 老年(66+))、hours per week (兼職(0-25), 全職 (25-40), 過度 (40-60), 極度(60+))、capital gain & capital loss (無(0), 低(介於 0 和 max 之間), 高(大於 max))。取 X_train 前 70%作為訓練用，後 30%作為驗證用。
來自 X_train 與 X_test 的特徵皆會由特徵標準化處理過。
按照在 X_train 時的順序，依序取為一個 batch (例如：1-3100、3101-6200)，每一回合 epoch 內，batch 訓練的順序是隨機取的。(batch_size: 3100 epoch: 500)
有實做 ada_grad 和 regularization。(regularization constant: 20)
準確率：0.85504 (kaggle)

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：若無特徵標準化，模型更新速度比較緩慢，跑 500 個 epoch，實作特徵標準化前後，在切出來的驗證集上，準確率分別是 0.81697/0.85791。觀察到無實作的情況，準確率有穩定上升的趨勢，故增加到 3000epoch，最後得到的準確率為 0.84277。比較兩者，可以發現有實作後，能夠增快模型更新效率，得到比較好的正確率。

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：藉由試誤法，調出的正規化常數 λ 為 20。實作正規化前後，在切出來的驗證集上，準確率分別為 0.85761/ 0.85791。拿到 kaggle 上測試，分別得到的準確率為 0.85369/0.85504。實作後，在兩種測試上，表現確實有好一點。

5.請討論你認為哪個 attribute 對結果影響最大？
答：針對幾項 attribute，看拿掉後，在切出來的驗證集上，對於準確率的大小有何影響作為判斷依據。然後為了相同比較基準，固定用 500 epoch。
完整的版本準確率：0.85791

第一項、遮 ada_drad，準確率 0.85822。(不過在 kaggle 上，正確率下降很多，還有加入 ada_drad，可以幫助模型收斂更好)
第二項、遮 regularization term，準確率 0.85761。
第三項、遮 feature normalization，準確率 0.81697。(主要影響因素：收斂速度)
增加 epoch 數目到 3000，可以讓準確率提高至 0.84277。
第四項、batch_size， 使 batch_size 等於訓練資料的筆數，準確率 0.83202。
第五項、訓練 batch 順序不隨機，準確率 0.85535。
隱藏項、不處理連續輸入特徵，準確率 0.84962。

從準確率來看，第四項的 batch_size 影響最大。