學號：B03901101 系級： 電機三 姓名：楊其昇

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：
將助教抽好的 feature（X_train）裡所有的 attribute 全部當成 Gaussian distribution 的模型下去算，當>50K 與 <=50K 的兩筆資料未共用同樣的 sigma 時，Kaggle 上 public 的準確率為 0.81327；但當兩者共用同樣的 sigma 時，Kaggle 上 public 的準確率進步為 0.84103。

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：
同樣將助教替我們抽好的 feature（X_train）裡所有的 attribute 當作參數，並在額外加入其中的前六項的連續資料的平方與立方項、從 train.csv 所取得的 education_num 的平方和立方項、以及 age 的立方與 sex 和 capital_gain 的立方的分別乘積，總共 122 項參數。訓練時的相關參數，除了 learning rate 設成 0.5，其餘一律初始值均設為 0，iteration 為 2000 次，更新參數使用 AdaGrad，並採取 full batch 的方式，最後訓練出來 Kaggle 上的準確率有 0.85700。

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：
不論是 generative model 或是 discriminative model 在此次作業中，我均有做特徵標準化。對於 generative model 而言，特徵標準化的好處在於可以避免某些原始數據較大或較小的 attribute 影響進而增加準確率；對於 discriminative model 更是要做特徵標準化，否則 logistic regression 中的 sigmoid function 會因為有些 attributes 的值太大而產生 overflow，甚至會直接導致最後的輸出結果爛掉，特徵標準化在這次作業中，尤其是 logistic regression，我認為是個必要的手段。

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：
這次作業由於參數比較多，所以有時候會有一點點的 overfitting 的現象產生，因此不同於作業一幾乎無感的正規化，這次作業會如果正規化的 lambda 調得好的話（大部分都是調到變爛），確實會有可能產生較佳的預測模型，但進步也只有幾乎無感的變化（相較於沒有做正規化，做了之後，8000 多筆測資也只多對兩三筆而已）。而且輸入的參數一改變，就要重調一次 lambda，考慮到調整 lambda 曠日費時且進步有限，加上每日上傳次數的限制，所以這次作業中並沒有選擇做正規化的處理。

5.請討論你認為哪個 attribute 對結果影響最大？
答：

根據上傳到 Kaggle 上的準確率判斷，基本上我認為年紀是最大的影響因素。一方面年紀為每個人一定都有而且為連續的 attribute，另一方面就常理而言，年紀和年收入確實存在著某種程度的正相關（二十歲左右年收入便大於 50K 畢竟是極少數）。因此在本次作業中，想加強某項 attribute 的表現，通常便會將其和年紀相乘並再搭配次方的變化。