學號：B04705014 系級： 資管二 姓名：梁淯程

1.請說明你實作的 generative model, 其訓練方式和準確率為何？
答：
不用訓練。先把 training data 分成答案是 class 1 跟 2 的兩份，然後算出個別的 feature 平均以及一個加權後的共變異數，就可以求出兩個可以高機率 generate 出那兩份 training data 的 gaussian distribution，接著用貝氏定理就可以直接算出 testing data 屬於哪個 class 的機率比較高，得到答案。準確率是 0.82975。

2.請說明你實作的 discriminative model, 其訓練方式和準確率為何？
答：
我用的 feature 是助教提供的 X_train 裡的所有欄位，加上把前六欄個別取 0.5,2,3 次。使用的方法是 logistic regression(optimizer 是 adam，epoch 是 2000)，並用 cross validation 檢測各種 features 的取法。準確率是 0.85516。

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：
在第 2 題的實作方法裡已經包含了標準化(先算出次方項再全部一起標準化)，把標準化拿掉以後的準確率是 0.80386。

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：
正規化中的 lambda 太大的話會讓 training 的效果變差，太小的話則可能會有 overfit，不過我在第 2 題的實作方法中加入 lambda 為 1e-4 和 1e-5 的正規化，發現對增加準確度沒什麼幫助(1e-4 -> 0.85491 1e-5 -> 0.85332)

5.請討論你認為哪個 attribute 對結果影響最大？
用 cross validation 檢測後發現拿掉 capital gain 以後對準確度的影響最大，不過其實就直觀而言收入跟 capital gain 本來就是很像的概念，因此相關程度很高也很合理。