學號：B03705006 系級： 資管三 姓名：侯舜元

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：
取 workclass, marital-status, race, occupation, native-country, education 當作 feature，機率分布使用 bernoulli 分布。透過嘗試訓練後將 cut off value 設為 0.45。如果 P(C1|x) > 0.45，為 class 1(<50K)，反之亦然。準確率為約 0.8256196062774485。

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：
除了 fnlwgt 外取全部的 feature，連續的資料有經過標準化。使用 gradient descent 去 minimize cross entropy。Feature vector size 105，Iterate 1000 次。隨機取 90%的 training data 下去 train，最佳準確率為約 0.8534443045361014，普遍上皆比 generative model 好。

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：
對於連續的 feature 皆標準化，其準確率為約 0.8534443045361014。取同樣的 feature 但不作標準化的準確率約為 0.830963422499309。可以看出 feature normalization 對於準確率有顯著的幫助。

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：

$\lambda = 0, \text{Accuracy} = 0.853444$	$\lambda = 0.5, \text{Accuracy} = 0.85154018$	$\lambda = 1, \text{Accuracy} = 8492368170$

實作 logistic regression 的正規化後，λ愈高準確率反而下降，看來正規化對於模型準確率沒有太大的幫助。

5.請討論你認為哪個 attribute 對結果影響最大？
除了 fnlwgt 外取全部的 feature，連續的資料有經過標準化，經過 training 得到的最佳準確率為約 0.8534443045361014。
試著一次拿掉一個重要的 feature
移除 education : 0.8413746506556924
移除 workclass: 0.8504959921378336
移除 marital-status : 0.8524308221491969
移除 occupation: 0.8463806394152513
移除 age : 0.8503117226129419

移除 hours-per-week: 0.8500967415005681
可以看除移除 education 這個 attribute 之後，準確率下降的幅度最大，可以算是最有影像力的 feature 之一。另外 occupation 也是一個用來判斷的很重要的指標。值得一提的是，原先我陷入無法進步的瓶頸，後來將一些看似很難最為判斷標準的連續 feature(例如：hours per week)納入 model 考慮後，準確率就有再上升不少。