學號：B04902044 系級： 資工二 姓名：朱柏澄

1.請說明你實作的generative model，其訓練方式和準確率為何？
答：我用Gaussian Distribution作為模型，並假設兩個class的Gaussian distribution有相同的covariance matrix。帶入公式後得到w, b，即為我的模型。
其準確率在public data上為0.76032。

2.請說明你實作的discriminative model，其訓練方式和準確率為何？
答：我將非binary的feature都做了normalization，並且加上二次方和三次方項，其餘binary的保持不變。使用adagrad，learning rate設為0.02，加上L2 regularization, lambda = 0.001，跑了9001 epoch。
其準確率在public data上為0.85786。

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：未做標準化的準確率：0.79115；有做標準化的準確率：0.85786
我將非binary的feature做標準化後，準確率有大幅提升。我認為標準化對於我的模型有增加準確率的效果，因為那些非binary的feature的數值的範圍太大了，function會被數字很大的feature dominate，做了標準化使數值都在差不多的範圍，可以避免這樣的狀況。

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：未做正規化的準確率：0.85737；有做正規化的準確率：0.85786
我加上L2 regularization, lambda = 0.001，結果和沒有做時相去不大。我認為正規化對於我的模型並沒有太大的影響，我將未作正規化的模型的結果印出來，發現w的範圍差不多都在1e-1~1e1，就已經蠻小的了，因此做了正規化後才沒有明顯的改變。原因是因為我有做了標準化，並且設initial weight為0，learning rate = 0.02，所以每次更新w都不會跑太遠，都在0的附近。

5.請討論你認為哪個attribute對結果影響最大？
一開始我認為助教給的檔案中，one-hot的feature有較大的影響，因為它們幾乎佔了feature的大部分。我在網路上找了處理categorial variable的其他方法，並發現binary encoding有著不錯的效果(https://goo.gl/1KuCnD)，而且還可以大幅降低feature的維度。但我實作binary encoding後，發現結果並沒有比較好，不論是在generative model 或 discriminative model 上都沒有變好，所以

我認為那些binary features對結果影響不大。最後我一次移除一個非binary的feature，觀察validation的準確率，發現移除”age”這項feature後的準確率最低（見下表），因此我認為**”age”**對結果影響最大。

Removed feature	Accuracy
age	**0.854431**
fnlwgt	0.85854
capital_gain	0.85793
capital_loss	0.85799
hours_per_week	0.85707