學號：B03705012　系級：資管三　　姓名：張晉華

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：
model 是取用 train.csv 裡的全部資料，保留連續的資料，然後針對多選的資料去算條件機率來代表並假設它是連續資料(ex:以 sex 舉例，對於選項 Male 去算機率 P(>50K|Male)來代表)，最後在全部標準化(使平均數=0 且標準差=1)就是我的 model，在假設為高斯分布作 Probabilistic Generative Model，Kaggle 上 public 分數為 0.83993，private 分數為 0.84019

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：
Discriminative model 的資料跟上一題(Generative Model)的資料一樣，但多加上了各個特徵資料的平方項，訓練方式為 logistic regression。Learn_rate 初始值為 0.5，若遇到 Cross entropy 值變大則將 Learn_rate 除 2，當兩次相差不到 10 時視為學習完成，Kaggle 上 public 分數為 0.85455，private 分數為 0.85542

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：
以 Probabilistic Generative Model 測試
未標準化的結果：Kaggle 分數(public,private) = (0.76855,0.77718)
標準化的結果：Kaggle 分數(public,private) = (0.83993,0.84019)
可以得知特徵標準化會有助於提高準確率

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：
在 learn_rate 初始值為 1.0，相差值不到 10 就停止，以 logistic regression 測試

lambda	Kaggle 分數(public,private)
0(沒實作 regularization)	(0.85455,0.85542)
10^{-9}	(0.85455,0.85542)
10^{-8}	(0.85516,0.85456)
10^{-7}	(0.85455,0.85542)
10^{-6}	(0.85455,0.85542)

可以得知，適當的 lambda 可以提升準確率，但影響似乎不大
5.請討論你認為哪個 attribute 對結果影響最大？

因為有做特徵標準化平均數與標準差各個特徵都相同，因此 w 的絕對值越大則可一定程度代表這個特徵對結果的影響性。

w('age','workclass','fnlwgt','education','education_num','marital_status','occupation','relationship','race','sex','capital_gain','capital_loss','hours_per_week','native_country','age 平方','workclass 平方','fnlwgt 平方','education 平方','education_num 平方','marital_status 平方','occupation 平方','relationship 平方','race 平方','sex 平方','capital_gain 平方','capital_loss 平方','hours_per_week 平方','native_country 平方')

= [0.79573828,-0.13671305,0.09658272,0.13474209,0.53602546,0.45081513,0.52295407,0.74295322,-0.45472065,-6.94106434,2.38666339,-0.0084049,0.36339203,0.10941222,-0.45405983, 0.05818218,-0.0166682,-0.00330726,-0.0003663,0.12324745,-0.10567612,-0.02382353,-0.21525751,-9.67087668,-0.1369214,0.05978496,-0.08683489,-0.01665224]

由上可知，可以推論”sex”跟”capital_gain”的影響最大