學號：B04902045 系級：資工二 姓名：孫凡耘

1.**請說明你實作的generative model，其訓練方式和準確率為何？**

我使用老師投影片中的方法，並使用助教給的X_train與Y_train做為training data. 我假設C1, C2機率的分布是一個multivariate gaussian distribution, 並實作老師投影片中的公式（假設兩個分布的covariance matrix是一樣的）。

主要使用公式：

P(C1 | x) = P(x | C1)*P(C1) / P(x)

P(x) = P(x | C1)*P(C1) + P(x | C2)*P(C2)

另外有一個參數可以tune，也就是gaussian distribution的公式其實是機率密度，故真正的機率可以視作再取個alpha次方。最後我去alpha = 1.55得到最好的結果（I wrote a shell script here to tune this parameters.）。

Alpha = 1, Accuracy = 0.83342

Alpha = 1.55, Accuracy = 0.8441

2.**請說明你實作的discriminative model，其訓練方式和準確率為何？**

I used logistic regression with decreasing learning rate and regularization. (No mini-batch)

Coss function: cross entropy

Learning-rate: 1e-9

Decrease rate: 0.999 after 100 iterations

Iterations: 100000

lambda(regularization): 0.001

用所有X_train裡面的feature

Accuracy: about 0.853(no matter how I tune those parameters)

後來我多加了五個feature，即所有continuous terms 的 ln(1 + x) 項

Accuracy: 0.8557(on public data set)

Code segment:

```
( xDf and testDf are dataframes in panda )
continuous = ["age", "fnlwgt", "capital_gain", "capital_loss", "hours_per_week"]
for ff in continuous:
  xDf[ff+"*"] = np.log( 1+xDf[ff] )
  testDf[ff+"*"] = np.log( 1+testDf[ff] )
```

3.**請實作feature normalization並討論其對於你的模型準確率的影響。**

因為我只用助教幫我們處理過的X_train and Y_train，我只對continuous terms做normalization（我試過把所有feature都normalize表現都會變很差，而且直覺上normalize這種discrete的data沒有意義，雖然不確定為什麼會變差而且變差蠻多的。）.

Code segment:

```
continuous = ["age", "fnlwgt", "capital_gain", "capital_loss", "hours_per_week"]
for ff in continuous:
    xDf[ff] = (xDf[ff] - xDf[ff].mean())/xDf[ff].std()
    testDf[ff] = (testDf[ff] - testDf[ff].mean())/testDf[ff].std()
```

Logistic Regression(using only features in X_test) :

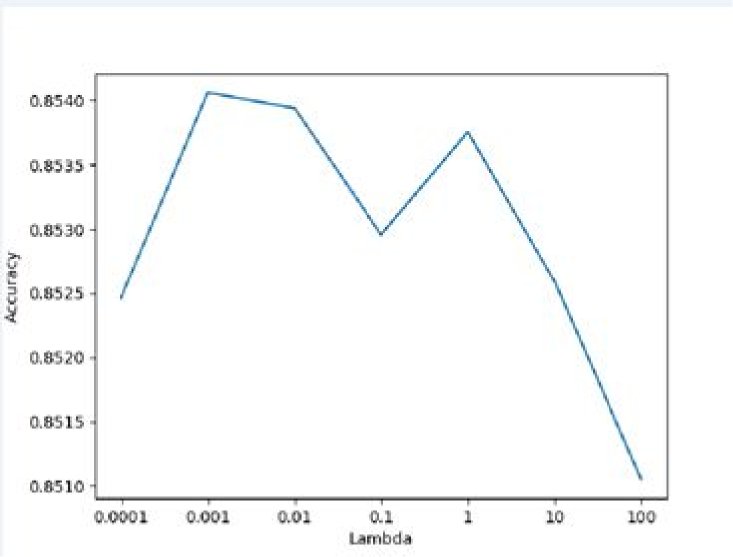
Accuracy without normalization: 0.8514
Accuracy with normalization: 0.8524

Generative

Accuracy without normalization: 0.8441
Accuracy with normalization: 0.8443

Generally, using normalization on continuous terms do boost the result.。 兩種模型中又已logistic regression的影響比較大。

4. 請實作logistic regression的regularization並討論其對於你的模型準確率的影響



I tried 7 different values for lambda. 結果如圖。不太確定為什麼會有這樣的上下起伏，但我們可以確定的是lambda的值對regression的結果確實會有影響（儘管其實差異的範圍沒有很大），而且大體上來說會有一個”高峰”。

5.**請討論你認為哪個attribute對結果影響最大？**

我跑了一個script，每次drop一個feature，並用generative的model看看Accuracy如何。

Dropped features	Accuracy	Dropped features	Accuracy
age	0.84300	wokclasses	0.84270
fnlwgt	0.84337	countries	0.84398
sex	0.84288	jobs	0.83704
capital_gain	0.83711	martial_status	0.84411
capital_loss	0.84177	races	0.84398
hour_per_week	0.84417	relationship	0.84221

From the above table, I think jobs can be considered as the most important attribute.