學號：B04902112 系級： 資工二 姓名：張凱捷

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

使用provided training data的每一個feature和年紀的三次方，使用Gaussian Distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

- 計算每一比data <50K的機率是否大於0.5，來決定答案是0 or 1

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

準確度：training data: 0.84493　　public testing data: 0.84349

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

使用provided training data的每一個feature(fnlwgt除外)和連續資料(data[0]~data[5])的0.5~9次方
使用logistic regression，learing rate: 1.0， 1600 epochs

- 透過validation發現traning data的準確度和validation的準確度差不多，於是我便嘗試加入各種feature儘量使traning data的準確度提高，先不考慮overfit。

準確度：training data: 0.86053　　public testing data: 0.86002

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。

```
for i in range(FFF):
    div.append(numpy.average(train_in[:,[i]]))
    std.append(numpy.std(train_in[:,[i]], ddof=1))
for i in range(FFF):
    train_in[:,[i]] = (train_in[:,[i]] - div[i]) / std[i]
    vi[:,[i]] = (vi[:,[i]] - div[i]) / std[i]
```

(a) 以我的logistic regression(第2題)來說，若沒有標準化，則高次方的feature有可能會overflow，或是各種誤差導致準確率下降。
normalize　　training data準確度： 0.86053
without normalize　　training data準確度： 0.73311 (爆炸)

(b) 以我的generative model(第1題)來說，反而是沒有標準化準確度比較高
normalize　　training data準確度： 0.80421　　public準確度： 0.80049
without normalize　　training data準確度： 0.84493　　public準確度： 0.84349

(c) 若拿原始Data並扣掉較大的feature fnlwg做logistic regression，結果發現最後的結果相似，但是標準化可以加快訓練的速度。

epochs	1000	2000	3000	4000	5000	6000	7000	8000	9000
normalize	0.8462	0.8504	0.8512	0.8520	0.8524	0.8528	0.8530	0.8532	0.8533
original	0.8533	0.8533	0.8533	0.8533	0.8533	0.8533	0.8533	0.8533	0.8533

- 由(a)(b)(c)可得知，有沒有標準化沒有絕對的好壞，可能會有精度問題，或是使原始資料某些性質不見

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。

- 將loss function加上 x^2項，使model參數不會太大

```
Fwb = sig(train_in.dot(mod))
grad = numpy.transpose(train in).dot(Fwb - train out) + 2 * lamda * mod
```

使用第2題的model來比較lamda

lamda	0	0.01	0.1	1
training	0.86053	0.86020	0.86078	0.86056
public	0.86002	0.85897	0.85860	0.85835

- 由上表可知正規化對於我這題的model的準確度並沒有幫助，有可能是因為本題traning data和testing data相似，noise不多

5.請討論你認為哪個attribute對結果影響最大？

我透過加一個，少一個feature的方式，來決定要不要選某一個feature
以training data準確度來看

all feature:	0.85332
no age:	0.85166
no fnlwgt:	0.85243
no sex:	0.85295
no capital_gain:	0.83824
no capital_loss:	0.85120
no hours_per_week:	0.85132

- 對結果影響最大的feature為 capital gain