1.請比較你實作的generative model、logistic regression的準確率，何者較佳？
A:

Generative model:
把106維資料全當作 feature ，把資料分為兩類，分別計算 u 與共用的 sigma，並照投影片公式完成gaussian Distribution，上傳 kaggle 就超過 simple baseline了。

Logistic model:

同樣把106維資料全當作 feature ，但是有對對資料做 mix-max scaling，並採用 Adagrad 調整 learning rate.

Iteration	learning rate	training accuracy	public accuracy
3500	0.5	85.5680655067	0.85393

在我的實驗中，logistic model 的確能得到比較好的成績，我想是 sample 數目不算少，所以 logistic model 可以訓練得比較好。

2.請說明你實作的best model，其訓練方式和準確率為何？
A:

主要差別都在feature取法，我分別以 [0, 1, … , 105] 表示106 dim，設定參數為 learning rate = 0.7，iteration = 3000

```
a = X_train[:,1:]
b = X_train[:,[0,1,3,4,5]]**2
c = X_train[:,[0,1,3,4,5]]**3
d = X_train[:,[0,1,3,4,5]]**4
e = np.log(X_train[:, [3]] + 1e-100)), axis=1)
```

我主要把 feature 分成這幾類，組合起來訓練。

feature	training accuracy	public accuracy
a + b + c	85.8751279427	0.85810
a + b + c + e	86.5233708632	0.86142
a + b + c + d + e	86.5916069601	0.86179
a + b + c + d + e改為[:, [0,3]]	86.8645513477	0.86203

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。

A:

若是沒有做 feature normalization，training 效果會差很多，因為 feature 大部分主要是 0 與 1 但是前六項卻是相對較大的數字，如果不做 feature scaling ，會收斂過慢。

feature normalization	Iteration	learning rate	training accuracy	public accuracy
min-max scaling	3000	0.5	85.5680655067	0.85285
none	3000	0.5	78.2668031389	0.79742

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。

A:

在這題中我設定參數為 learning rate = 0.7，iteration = 1000, feature取 106 dim

Lamda	1000	100	10	0.1
training accuracy	84.7833503924	85.4998294098	85.5680655067	85.5680655067

做regularization後，對模型準確率的影響並不大，我想是模型並沒有太過複雜(Feature 的選擇，哪些要做 n 次的選擇還不至於使模型變得複雜)，所以並不會不會訓練資料過度，產生overfitting。

5.請討論你認為哪個attribute對結果影響最大？

A: 我認為 **capital gain** 的影響最大 ，我加上 $\log(capital\ gain)$ 後，模型收斂速度快很多，再加上將連續性的 feature 做n次方在 public set 上能取得較好的成績。