學號：B04701232 系級： 工管三 姓名：陳柔安

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

model	generative model	logistic regression
accuracy	0.8525889073152755	0.8571955039616731

兩者使用的feature皆相同（同best model），並把全部維度都做feature normalization，由結果看來，logistic regression的結果稍好於generative model。

2.請說明你實作的best model，其訓練方式和準確率為何？
答：
training feature 取用助教抽好的data中除了fnlwgt以外的全部項目，並將age, capital_gain, capital_loss, hours_per_week項目取二次項與三次項加入feature中，再加入bias。使用logistic regression的方法training，train的過程中有使用adagrad，並做正規化 lambda=1，採用full batch size，並iterate 100000次。
此模型的準確率為：
private:0.85628(kaggle)
public:0.85823(kaggle)
total:0.857257(本機測試)

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：
下表為各種feature下是否有做feature normalization的accuracy，logistic採用best model，唯iteration改為10000次。

	有標準化	無標準化
logistic只有一次項	0.8518518518518519	0.7984767520422579
logistic加入二次項	0.8560899207665377	0.8253792764572201
logistic加入二次項與三次項	0.8574411891161476	0.724034150236472
generative	0.8525889073152755	0.23622627602727106

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：

下表為各種feature下是否有做regularization的training and testing accuracy，logistic model採用best model，唯iteration改為10000次。

lambda	training	testing
0	0.8578667731335032	0.8574411891161476
0.01	0.8578360615460213	0.8574411891161476
0.1	0.8577746383710574	0.8574411891161476
1	0.8574368109087559	0.857379767827529
10	0.8572525413838641	0.8563356059210122
100	0.855256288197537	0.8557828143234445

根據測試結果lambda愈大training and testing accuracy皆會微幅升高，可推測此模型下是否有做regulization對於準確率影響不大，且以不做之結果為佳。

5.請討論你認為哪個attribute對結果影響最大？
將training data各個參數與其label討論相關性，若為連續變數算出相關係數。若為非連續變數，以one hot encoding方式算相關係數取平均，結果如下：
age:0.234037
fnlwgt:-0.009463
sex:0.215980
capital_gain:0.223329
capital_loss:0.150526
hours_per_week:0.229689
work_class:0.011125
education:0.002659
marital_status:-0.024282
occupation:-0.008766
relationship:-0.019884
race:-0.010774
native_country:-0.002562
由結果可推測出age, sex, capital_gain, capital_loss以及hours_per_week這幾個attribute與預測結果相關性較高，間接推測其對結果影響較大。