學號：R06944031　系級：　網媒碩一　姓名：林蔚廷

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

答：

#Generative model score:

private:0.84227, public:0.84533

#Logistic model score:

without regularization

learning rate = 1e-5

iteration = 10000

private:0.85112, public: 0.85454

在我的 model 中 logistic regression 的準確率，在 public 和 private 下都表現得比較好。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我取了所有的 Feature，除了編號 78: Holand-Netherlands 之外，因為此 feature 會產生除以零的情況，並且加入 2 次式和 3 次式，用 1e-5 的 Learning Rate 學習，執行 10000 次 Iterations，Regularization 的 lambda 取 1

準確率:

在切半的 Cross validation 中平均準確度: 0.8559627300298942

Kaggle 上的分數: private:0.85665, public:0.85933

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：

因為我最好的模型採取三次式，執行時間很慢，且表現只比二次式好一點點，所以這邊以一次加二次式來實驗。

With normalization:

Learning rate: 1e-5

Cost: 降到 0.315795

準確度:　private:0.85112, public:0.85454

Without normalization:

Learning rate: 1e-15

Cost: 降到 0.542787

準確度:　private: 0.78835, public: 0.79299

加入特徵標準化，可以大幅度的增加模型的準確率

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：

模型: 一次與二次，有 normalization，Learning rate: 1e-5 Iteration: 5000

Lambda	**Private score**	**Public score**	**Average**
0	0.85554	0.85835	0.85694
0.1	0.85566	0.85835	0.85700
1	0.85591	0.85847	0.85719
10	0.85456	0.85835	0.85646
100	0.85186	0.85663	0.85425

Lambda = 1 表現最好

5.請討論你認為哪個 attribute 對結果影響最大？

在一次式模型下觀察，learning rate = 1e-5，iteration = 2000

兩個 set，cross validation，分別刪除不同的 attribute 觀察準確度變化

Feature	Accuracy1	Accuracy2	Avg acc
All data	0.8511056511056511	0.8493335790184878	0.8502196150620694

Delete feature	Accuracy1	Accuracy2	Avg acc
0	0.8515970515970516	0.8486579448436828	0.8501274982203673
1	0.8511670761670762	0.8486579448436828	0.8499125105053795
2	0.8514127764127765	0.8502548983477674	0.850833837380272
3	0.8385135135135136	0.833978256863829	0.8362458851886713
4	0.8512285012285012	0.8462625145875561	0.8487455079080286
5	0.8507565123572101	0.8486579448436828	0.8493289724218414
6~14	0.8506142506142507	0.8468767274737424	0.8487454890439965
15~30	0.8439189189189189	0.8401203857256925	0.8420196523223057
31~37	0.8519041769041769	0.8501934770591487	0.8510488269816627
38~52	0.8482800982800983	0.8433757140224802	0.8458279061512892
53~58	0.8519041769041769	0.8487193661323015	0.8503117715182391
59~63	0.8508599508599508	0.8501320557705301	0.8504960033152404
64~105	0.8515356265356265	0.8501934770591487	0.8508645517973876

再刪除不同 Feature 下可以觀察到，刪除 index = 3 的 feature(capital_gain)，準確率下降最多，我認為此 feature 對我的模型結果影響最大，若只取此 feature 做 training 可以在 validation set 得到平均.79795 的準確度