Machine Learning Homework 2 - Income Prediction

學號：B04901061 系級：電機三 姓名：蔡忠紘

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

答：

	Private Score	Public Score	Average Score
generative model	0.84645	0.84889	0.84767
logistic regression	0.85579	0.85798	0.85689

不論是private或是public，使用logistic regression的準確率都較好。

2.請說明你實作的best model，其訓練方式和準確率為何？

答：

實作準確率最高的是使用Scikit-learn的套件，但我認為手刻logistic regression花費較多心思，因此在這一併附上訓練方式。

使用logistic regression產生最好結果的參數設定為：

（1）加入age、capital_loss的二次項

（2）Normalization

（3）No Regularization

（4）Iteration = 2000、Learning Rate = 1.1

Accuracy：Private Score = 0.85579，Public Score = 0.85798

另外，使用Scikit-learn的RandomForestClassifier，參數設定為：

（1）n_estimators = 450

（2）max_depth = 13

Accuracy：Private Score = 0.86156，Public Score = 0.86461

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：

測試條件：Adagrad、No Regularization、Iteration: 2000

	Private Score	Public Score	Average Score
without normalization	0.65999	0.66683	0.66341
normalization	0.85579	0.85798	0.85689

這次的feature數量級相差很大，Normalization的效果相當顯著！除了可以加速training外，做Adagrad時也比較不容易卡在 saddle point 或 local min。

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：

測試條件：Adagrad、Normalization、Iteration：2000

λ	0.001	0.01	0.1	0.5	1	10	100
Private Score	0.85579	0.85579	0.85554	0.85530	0.85431	0.85259	0.85050

從實驗中可得觀察到，regularization的效果並不顯著，但在λ太大時，會有反效果，尤其在λ大於1時Accuracy降得更劇烈。

5.請討論你認為哪個attribute對結果影響最大？

答：

從weight的分佈來看，似乎是age和capital_gain的影響較大，且capital_gain明顯和年收入是否大於50K呈正向關係，簡單瀏覽train.csv中的資料，也可發現capital_gain欄位非0的受試者年收大多大於50K。

推測是因為，收入要高到一定程度才有餘裕投資，才會有capital_gain。