HW2 Report

學號：R06942074 系級： 電信所碩一 姓名：李宇哲

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

從整理過的下表可以發現，Generative model 和 Logistic regression 在準確率的表現上有明顯的差異，例如 Generative mode 不論在 Private date set 或者 Public data set 的表現都比較差，甚至在自己所切的 validation set 也不好。

	Private score	Public score	Validation score
Generative probability model	0.84191	0.84594	0.837
Logistic Regression model	0.84866	0.85466	0.855

(Validation set size = 0.1 * total data)

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

因為這次的 best model 可以使用 xgboost 套件，所以我使用 xgboost 裡面的 XGBClassifier () 作為這次的 best model。xgboost 的訓練方式主要是改進吃 boosting tree 及考慮二次微分去 optimize weight，我的每棵樹最大深度只有 5，然後有 300 棵樹，learning rate 給 0.05。

準確率：

	Private score	Public score	Validation score
Xgboost	0.87311	0.87800	0.877

這裡可以發現準確率明顯都比 generative model、logistic regression 好

Refernce:

Tianqi Chen, Carlos Guestrin. “*XGBoost: A Scalable Tree Boosting System.*” KDD'16

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：
我使用的特徵標準化如下：

$$x' = \frac{x - \bar{x}}{\sigma}$$

從下表可以發現，如果沒有使用特徵標準化，會大幅降低 Logistic Regression model 的準確率，推測可能跟第一個 feature 數值過大有關。

	Private score	Public score	Validation score
Logistic Regression model (No normalization)	0.79302	0.79778	0.81
Logistic Regression model	0.84866	0.85466	0.855

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：
由下表可以發現，加上 regularization 似乎沒有明顯的增加準確率，也就是說 model 在訓練時沒有嚴重的 overfitting 現象。

	Private score	Public score	Validation score
Logistic Regression model (No regularization)	0.85087	0.85417	0.853
Logistic Regression model	0.84866	0.85466	0.855

5.請討論你認為哪個 attribute 對結果影響最大？

從實驗結果跟比較來看，我覺得特徵標準化(feature normalization)是影響最大的原因，再來我覺得這次的作業有一個影響更大的問題就是 training data imbalance（3:1），如果要 train 一個 logistic regression，imbalanced labelled data 會使 model 產生 gradient dilution 的現象，在 MLP 的訓練或是其他跟 DL 有關的 model：CNN 都會因為 imbalance data 的問題使 model 在 training 的時候沒辦法接近 99% 的 training acc。通常會使用 up sampling 或者 incremental learning 去解決。