學號：B03902013 系級： 資工四 姓名：吳克駿

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

答：

generative model: public:0.84496 private:0.84338 average:0.84417

logistic regression: public:0.85466 private 0.85063 average:0.85265

Logistic Regression model 的準確率較佳

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

全取 X_train 中 106 維的資料，利用 xgboost 套件中 Classifier 中的 logistic regression 方法，並調整其 maximum tree depth 為 7，可得準確率

public:0.87874 private:0.87311 average:0.87593

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：

對於 Normalization 來說，我原先做的比較屬於 Rescaling，但應助教範例為 Standardization，因此後來我兩者都有做。

對於兩個模型的準確率來說

	Basic	*Rescaling*	*Standardization*
Generative	0.84350	0.84368	0.84417
Logistic	0.79516	0.85265	0.83219

在 Generative model 中，三者的準確率差不多，但對於我實作出的 Logistic Model 來說，三者就有一段差距。做 Normalization 可以使數據較大的 data，例如像 fnlwgt，範圍限制在[0,1]之間，使其較容易被計算且所小誤差；而 Rescaling 較 Standardizaion 好的原因，我認為是因為在 one-hot encoding 中，只有 0 與 1 的差別，因此 Rescaling 時會保留其 0,1 的差距，但 Standardization 卻會縮小其差距，模糊化兩者的差別，也因此做出來的準確率稍微較低。

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：

實作 L2-regularization 的方法為原本的 loss function 後面再加上$\lambda w^t w$，取小的 λ 會使整個曲線平滑一點。我取的 λ 為 1e-5。

實作出來的結果準確率 0.85253，跟原本做出來的 0.85265 相去不遠且未進步，原因應該是此次的模型並不複雜，因此正規化做出來的成果進步有限；也可能因為我兩次所做的 epoch 次數一致，可能需要跟長時間的訓練才能得到較好的結果。

5.請討論你認為哪個 attribute 對結果影響最大？

我將所有 arribute 都分別放入 generative model 裡面實作，去看看哪個 attribute 能得到最好的效果。

首先，因為某些 attribute 做出來答案都是 0，雖然其結果數字可能比某些結果高，但由於毫無參考價值，因此把這些剃除。其中有 fnlwgt、sex、race、native-country，這些應該皆為分組後每個細項 0 仍遠大於 1 的分布。

再來看其他的 attribute，做出來的結果為下

age	capital-gain	capital-loss	hours per week	workclass	education	marital status	occupation	Relationship
0.74504	0.77138	0.76506	0.75573	0.76702	0.78005	0.76340	0.75517	0.75966

最高的 attribute 為 education，代表學歷跟未來的收入相關性最大，也驗證了一般人直覺的想法：高學歷領高薪。