學號：R05943110 系級： 電子碩二 姓名：蕭 堯

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

答：**Logistic regression 準確率較佳**

Generative Accuracy: 0.844
Logistic Accuracy: 0.853

Logistic regression 沒有像 Generative model 作一些機率分布的假設，加上這次的 training data 有三萬多筆，算是蠻充裕的，所以我認為這也許就是 Logistic 的方法準確率比 Generative 高的原因。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

答： **Accuracy: 0.861**

我使用了 Keras 做一個 NN，我發現一層 Hidden layer 其實就算足夠了 ，太多層反而沒有更好的表現。

Input : 106 個參數中，我將所有 continuous 都 normalize，boolean feature(0 或 1)都剪掉 0.5(變成 0.5 或-0.5)，我發現 boolean feature 剪掉 0.5 會比 normalize 的處理好一點。
Hidden layer : 只有一層，100 個 neurons，採用 ReLu。
Dropout : 在 Hidden layer 後有 Dropout = 0.1。
Output : 輸出採用 Softmax。
Optimizer : RMSprop (lr = 0.0001)

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：

	Generative model	Logistic regression
w/o normalization	**0.842**	**0.784**
w/ normalization	**0.844**	**0.853**

Logistic regression 在沒有 normalization 的表現明顯比較差，我想可能是因為有些參數(ex: fnlwgt)相較於其他參數大上許多，這會使得在**沒 normalization 的情形下 SGD 在相同 learning rate 下更新 weights 時變得困難許多**。

Generative model 有沒有 normalization 似乎不太影響，我認為可能是因為 Generative model 不需要經過一步步更新參數的過程，因為已經有解析解了，**有無 normalization 對應到的都是同一組解析解，所以出來的結果也就比較不受影響。**

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：

λ	0.1	0.01	0.001	0.0001
Accuracy	0.802	0.832	0.853	0.842

λ= 0.001 時，Accuracy 有 maximum，這也是我這次上傳作業所使用的參數。

5. 請討論你認為哪個 attribute 對結果影響最大？

我把 Logistic 的前五大 Weights 所對應的 features 找了出來，如下圖所示：

	capital_gain	Wife	Married -civ-spouse	Masters	Prof-school
Weights	2.20	1.39	1.33	1.05	1.01

資本利得：雖然大多數人為 0，有資本利得的人似乎就很有機會 > 50K ，有很強的正相關。

家庭角色環境：已經結過婚，並且有穩定的家庭，也蠻讓人意外的是不可或缺的一項因素。

學歷：教育背景不可否認的也是很重要的一項因素，其中以 Masters 和 Prof-school 的學歷有比較高的正相關。