學號：R05222003 系級： 物理碩二 姓名：吳愷訢

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

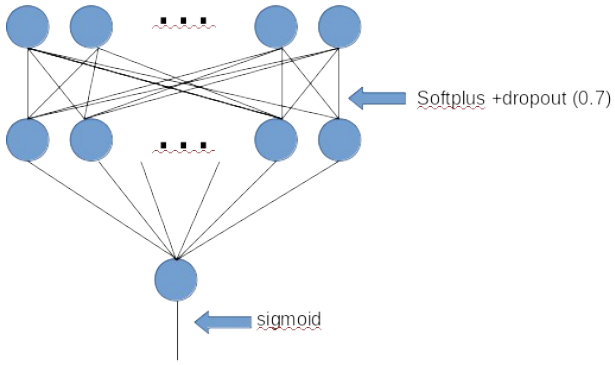
*以下以 self validation 與 Kaggle public score 來做分析. validation 部分爲 20% training data. Training set 使用 X_train ,Y_train 並移除 fnlwgt feature.

	validation	Kaggle public set score
Generative model	**0.84217**	**0.84471**
Logistic model	**0.849071**	**0.84987**

由以上結果我們發現 Logistic model 的結果較好.

2.請說明你實作的 **best model**，其訓練方式和準確率為何？
答：

在 best model 中,我使用 pytorch 實作類似 logistic model 的架構，但是在前面多加一層 layer, 作爲 feature map. 另外,因爲不知道哪些 feature 是重要的, 在考慮進可能有些 redundant feature, 在這裏使用 dropout. nn 架構如下：



在 training 方式使用 Stochastic gradient decent, Batch size = 256 . 一樣切 20% training set 作 validation. Accuracy 結果如下：

	validation	Kaggle public set score
Best model	**0.861020**	**0.85945**

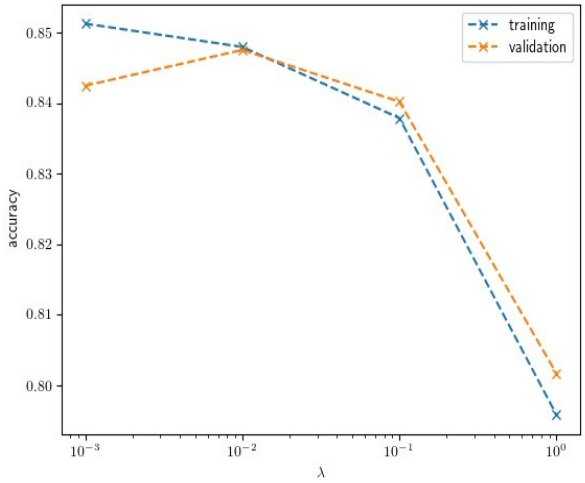
3.請實作輸入特徵標準化**(feature normalization)**，並討論其對於你的模型準確率的影響。
答：

	Raw	Norrnalization
Generative model	**0.83420**	**0.842800**
Logistic model	**0.83014**	**0.84987**

比較有無 normalization 的結果我們發現在做了 feature normalization 後準確率提高. 這是因爲 data distribution 對於各個類 feature 的散佈度不同,所以在 share covarience matrix 的架構下有作 normalization 較沒有作的準確率高.

---

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：



我們發現 validation accuracy 在 regulaization 增加至 0.01 時 accuracy 提升,接着下降, 意味着可能存在 overfitting. 故我們看到在做了 regularization 後 accuracy 提升.

5.請討論你認為哪個 **attribute** 對結果影響最大？

我認爲 workclass 跟 capital-gain /loss 影響應爲最大. 在實際上測試結果, 雖然無法 systematic 的分析各種 attribute 組合,但是我們可以發現 fnlwgt 應該是最沒有直覺相關的. 以下結果爲 simple logistic model 在有無 fnlwgt 下的結果：

	HAVE fnlwgt	WITHOUT fnlwgt
Logistic model	**0.84401**	**0.84602**

以上爲平均 5 次結果（每次 shuffle training set,並切 20% validation）. 我們可以看到在沒有 fnlwgt 下 accuracy 反而平均來說 performance 高了一點, 但是沒有差太多, 基本上是沒有直接強烈相關性的.