Homework 2 Report - Income Prediction

學號：R06942143 系級：電信一 姓名：籃聖皓

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

	Public	private
Generative model	0.84557	0.84191
Logistic regression	0.84877	0.84092

相較之下，logistic regression 的表現比較好
joint probability distribution 描繪出的結果，雖然也滿高的，不過這次的 data 用 conditional distribution 可以描繪得比較貼切。而也驗證了 discrimination model 表現通常會比 generative 來的好，不過 generative model 不用像 logistic regression 一樣訓練那麼久，算是其中之一的優點吧。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？
Best model 使用 svm 訓練
對於這樣的線性不可分的樣本，正好是 SVM 的長處
而懲罰因子的設定是 1.0
kernel 使用 rbf，再多維度的資料處理起來很合適
gamma = 1/# of feature
而準確度可以達到 0.87211

3. (1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)
Generative model 在還沒 scaling 之前的準確度是 0.84557
不過在 scaling 之後反倒是降低了準確度變成 0.76474
這邊可能的原因，覺得是因為 scaling 之後，讓樣本分布的資訊量變少，因此在判讀的時候降低了準確率。

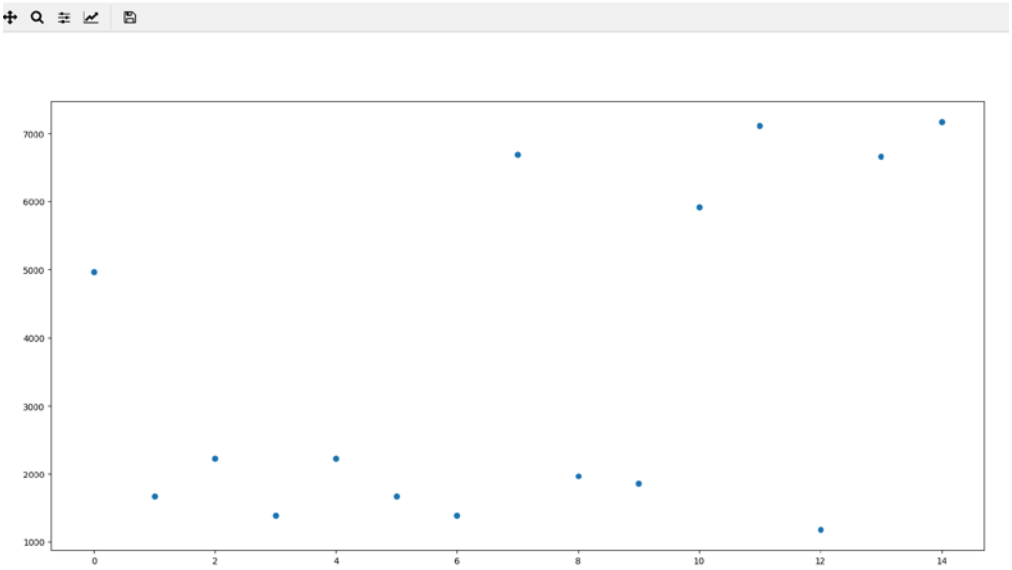
而 logistic regression 原本的準確度是，0.73820
在 scaling 之後準確度提高了成 0.84803
這邊推測的原因應該是，logistic regression 要描述的只是 conditional probability，而如果有些 feature 的數值遠遠大於其他 feature，這樣那些數值的影響就會特別的大，進而影響準確度。

4. (1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)
在使用了 regularization 之後，結果如下

lambda	score
0	0.84803
0.01	0.83857
1	0.76474

好像加入了 regularization 之後準確度反倒是降低了，這邊應該有兩個原因：
1. 沒有找到適合的 lambda parameter
2. 模型其實並沒有很複雜的 weight 項，加入 regularization 反而讓 gradientdescent 沒辦法到達 local minimum

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？
Age, fnlwgt, capital_gain, hours_per_week



是在收入大於 50K 的人之中，統計他們所擁有的特質，
最後看出擁有以上 feature 的人收入都會比較高
而刪掉這些 feature 之後的結果準確度

	只有 scaling	刪掉 attribute 較大 featuer
Public	0.84803	0.76474
private	0.84092	0.76280

而上面的 feature 也與現實情形符合，年齡、投資賺錢、每周工作，這些確實是影響收入較大的原因之一。