學號：B04901069 系級：電機四 姓名：林志皓

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

(The Accuracy below is evaluated over testing set on Kaggle)

Accuracy of Generative Model : 0.84631

Accuracy of Logistic regression: 0.85196

由上可知 ，logistic regression 得到的結果較佳

1. 請說明你實作的best model，其訓練方式和準確率為何？

我主要是在資料的處理多做一些事情。我把所有的continuous data 做 Discretization, 每筆資料都化為6維的one-hot-vector，分別代表"小於-2個標準差”、"介於-2到-1個標準差”、"介於-1到 0個標準差”、"介於0到1個標準差”、"介於1到2個標準差”、"大於2個標準差” 六個類別，而原本的連續資料還是保留，如此進行訓練。

過程中利用5-fold validation 驗證自己方法的準確性。

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

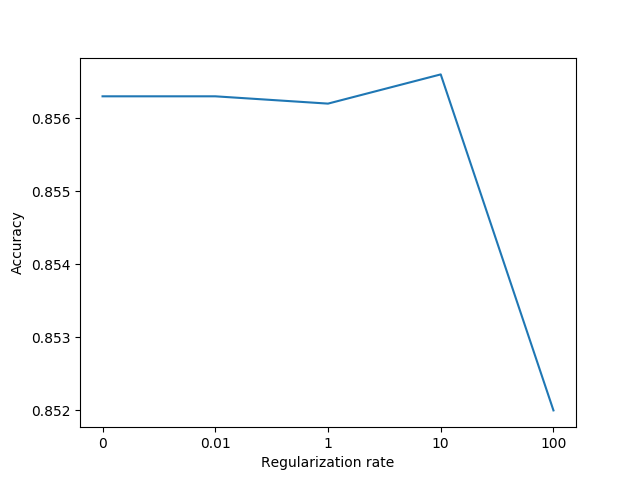
我實做的方式是針對每個類別continuous data，將其減去該類別平均，再除以該類別標準差。以下是實驗的結果：

Accuracy before feature normalization : 0.7958

Accuracy after feature normalization : 0.8520

可以看出feature normalization 讓準確率顯著提昇

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。



Regularization 用在model 複雜度太高，造成overfitting時所使用，讓分類的邊緣叫平滑，達到更好的generalization，在這次作業中，並沒有overfitting的狀況，因此做regularization並沒有顯著的效能提昇，也可能因為做太多而降低效能。由上圖所示，rate介於0.001 ~ 10 表現都差不多（10的時候好一點點），但當到達100時，就降低了表現。

1. 請討論你認為哪個attribute 對結果影響最大？

經過一些實驗，我認為capital\_gain這項資料影響滿大的，我對他做discretization 之後些顯著提昇。個人認為最多的還是對於所有continuous data進行feature normalization.