學號：r05922145 系級： 資工碩二 姓名：郁錦濤

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

答：

model	accuracy on public set	accuracy on private set
generative model	0.84203	0.84545
logistic regression	0.85050	0.85429

從兩種模型的在 public 和 private set 上準確率表現來看，logistic regression 的準確率更高。原因我認爲 generative model 是根據機率產生的模型，本來就是假設 data 滿足 Gaussian Distribution，所以會有偏差。

2.請說明你實作的 best model，其訓練方式和準確率爲何？

答：

model	**accuracy on public set**	**accuracy on private set**
Gradient Boosting Tree	0.86684	0.86842

我的 best model 採用了 Gradient Boosting Tree，它是一個支持 binary classification 的分類器，實作出來的準確率比較高。它使用 100 個單層的 decision tree 作爲一個弱學習器，通過組合來進行分類。

我主要用了 scikit-learn 套件中的 GradinentBoostingClassifier，通過喂入 X_train 和 Y_train，由程式自己生成 model，函式的參數均採用默認值，最後喂入 X_test，生成 Y_test。

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：

normalization	**accuracy on public set**	**accuracy on private set**
No	0.79766	0.79314
Yes	0.85331	0.85124

特徵標準化對於模型的準確率確實有很大的提高，並且加快了 training 的速度。

原本完整的 data 有 32562 筆，我將 data 先 shuffle，抽取 20%的 data 當做 validation set，因此實際用來 訓練的資料有 26000 多筆 ，而 feature 一樣是直接使用助教預先抽好的，訓練次數爲 3000 次，初始學習率爲 0.005，並且使用 adagrad。

可以看到在相同的訓練次數和初始學習率之下，feature normalization 能大幅提升準確 度和效能。另外，若沒有使用 feature normalization ，即使用了 adagrad，初始學習率也 必須好好選擇，否則容易在訓練過程中發散。

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：

regularization	**lamda**	**accuracy on public set**	**accuracy on private set**
No	0	0.85565	0.84977
L1	0.1	0.85466	0.85087
L1	0.01	0.85393	0.85201
L1	0.001	0.85331	0.85112
L2	0.1	0.77248	0.76992
L2	0.01	0.85589	0.84928
L2	0.001	0.85454	0.85112

我分別採用了 L1 和 L2 兩種正規化方法和不做正規化的情況進行比較，發現在添加正規化以後，對模型的準確率會有一定的提升，但是提升的幅度不大，有時候採用了不合適的 lamda 也會使準確率大幅下降。

因此我們得知，有正規化的模型確實能夠避免 overfitting 情況的產生，能夠在 private set 上的表現和 public 上的大體一致。

5.請討論你認爲哪個 attribute 對結果影響最大？

答：

Removed attribute	**accuracy on valid set**
none	0.851571
age	0.854232
fnlwgt	0.851824
sex	0.851713
capital_gain	**0.838174**
capital_loss	0.852032
hours_per_week	0.853721
workclass	0.852323
education	0.844002
marital_status	0.853290
occupation	0.847026
relationship	0.853120
race	0.853307
native country	0.851821

我試著每一次從 attribute 種剔除一列 feature，通過觀察我的 valid set 的 accuracy 來評估每個 attribute 對模型預測的影響。

從上表可以知道，capital_gain 這個 attribute 對結果影響最大。仔細分析發現符合，因爲 capital_gain 是從一些資本商品中獲得的收益，必然會直接影響一個人的年收入。