學號：B04901155 系級： 電機三 姓名：鄭閎

1.請比較你實作的generative model、logistic regression的準確率，何者較佳？
答：
以下使用的 training data 是助教經過 one-hot encoding 抽過feature，我再針對連續分佈的數值(age, fnlwgt, capital_gain, capital_loss, hours_per_week)做 standardization 後得到的資料。public test data 約有 8140筆，我自己切的 validation data 為 training data 的三分之一約有 12187 筆。overall accuracy 為 public 和 validation 的加權平均。

accuracy	public (8140)	validation (12187)	overall
logistic	0.85331	0.85111	0.85199
generative	0.83292	0.82822	0.83010

logistic regression準確率較佳。
generative model 有假設分佈是高斯，對模型做出假設造成模型彈性降低，在訓練資料較少時可以有比較好的表現。但是這次訓練資料夠多，所以用較少假設而彈性較好的 logistic regression 訓練出比較接近真實的函數。

2.請說明你實作的best model，其訓練方式和準確率為何？
答：
使用scikit-learn套件的random forest classicier。訓練方式是做出很多的decision tree，每棵樹使用的資料由原本的訓練資料集合隨機取子集合。各樹用子集合的資料訓練完要預測testing data的時候，每筆資料輸入各棵樹，最後由各棵樹產生的結果投票。票數最多的結果作為最後的結果。
訓練參數：clf = RandomForestClassifier(n_estimators=17, max_depth=18)
n_estimators是幾顆樹，max_depth是樹最大的高度。

accuracy	public (8140)	validation (12187)	overall
random forest clf	0.86388	0.86114	0.86224

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：
使用特徵標準化的結果如第1題，無特徵標準化如下：

accuracy	public	validation	overall
logistic	0.81228	0.79658	0.80287
generative	0.83292	0.82822	0.83010

logistic 的部份若沒有標準化會導致參數在空間中在某些軸上特別陡峭而其他特別平緩，因此除了 learning rate 要調小以外也比較難收斂。而generative model和沒有標準化的結果完全一樣。我猜測可能是因為 generative 是比較要兩個類別誰的 posterior

probability 高，而正規化雖然數值不同，但並不會改變資料之間誰高誰低的關係，因此結果相同。

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：

accuracy	public	validation	overall
λ=0 (no regularization)	0.85331	0.85111	0.85199
λ=0.001	0.85257	0.85101	0.85163
λ=0.01	0.85343	0.85120	0.85209
λ=0.1	0.85442	0.85120	0.85249
λ=1	0.85417	0.85083	0.85217
λ=10	0.85393	0.85064	0.85196
λ=100	0.85442	0.85074	0.85221
λ=10000	0.84864	0.84543	0.84672
λ=1000000	0.77739	0.76773	0.77160

正規化似乎對準確率幾乎沒有太好的影響，一直要到 λ=10000 才因為gradient太大導致收斂得不好，讓準確率掉了一點點。進一步測試λ=1000000結果準確率更差。
其實訓練資料的準確率並沒有比驗證資料(validation data)的準確率高太多，因此可判斷沒有overfittig的問題，可能是 logistic regression 並不複雜，不容易overfitting。而沒有 overfitting 的情況下其實用正規化自然沒有什麼效果。

5. 請討論你認為哪個attribute對結果影響最大？
答：
從 logistic regression 的 weight 來看，不同的 attribute 之間並沒有太大的明顯差異，將 attribute 一個一個拿掉後對準確率也影響不大。原本感覺 capital_gain, capital_loss 會比較有關係，因為有錢人才比較會有資本的收益損失。但是單單是這兩個 attribute 對最後預測的準確率也沒有太大的影響。