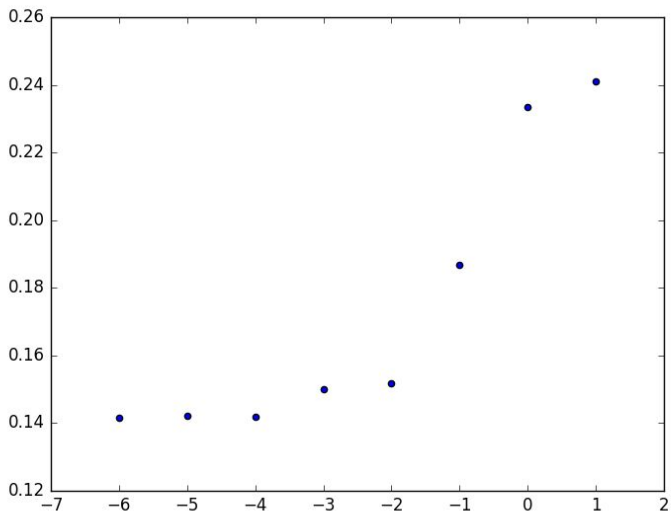
學號：B0b02014 系級： 生技三 姓名：張皓鈞

1.請說明你實作的generative model，其訓練方式和準確率為何？
答： 我利用了助教所抽取的one-hot-encoding的訓練資料，首先先把職業和教育程度作交集形成新的特徵，再去除掉不確定的特徵（有？的），再把連續值的特徵由低到高分成20類的類別特徵，假設訓練的資料是從高斯機率分佈中抽取出來的，並計算具有maximum likelihood的mu和covariance，計算方式為分別計算兩種class的平均和各項的共變異數，在根據class個數決定兩個class共用的共變異數矩陣。 利用所計算出來的模型在public data上的準確率為0.83771。

2.請說明你實作的discriminative model，其訓練方式和準確率為何？
答： 處理feature的方式和1.相同，以mini-batch size = 20的方式訓練logistic regression的模型，gradient descent的方式為adagrad，loss function的定義為cross entropy加上參數的L1正規化，我們選擇的λ為10^{-4}。 Traing Epoch=20的結果在public data上的準確率為0.85172。

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： 利用助教所抽取的one-hot-encoding的訓練資料，若未經過任何標準化和前處理，直接拿來訓練logistic regression的模型，其中模型的參數均和2.相同，在public data的預測準確率為0.66781。 而將連續值的特徵經過標準化後，在public data的預測準確率為0.67187。

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： 在此討論的為L1正規化，λ的範圍為10^{-6}，10^{-5}，10^{-3}，10^{-2}，10^{-1}，1和10，圖中x軸為log10(λ)，y軸為隨機取樣1/10的訓練資料作為validation set所測試的錯誤率。模型的參數和2.相同，處理feature的方式也和2.相同。比較圖如下：



可以看到validation set的錯誤率隨著λ的增加而增加，其中我選擇validation set錯誤率最低的$\lambda=10^{-6}$進行測試，在public data上的預測準確率為0.85258。

5.請討論你認為哪個attribute對結果影響最大？
答： 我認為「capital gain」和「教育程度是否為preschool」這兩個attribute影響最大，因為從訓練後的參數中可以發現對應到capital gain的weight數值最大，而對應到Preschool的weight數值最小。同時也符合我們常識中的兩個事實：投資賺錢較多的人收入會比較多以及受過高等教育的人收入會比未受高等教育的人還多。