

Homework 2 Report - Income Prediction

學號：r06725028 系級：資管碩一 姓名：黃于真

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

使用相同特徵，都有做特徵標準化，在public/private set上結果如下表，logistic regression的準確率比generative model多了一成多左右，算是差異很大，而且依據實際跑的情況，logistic regression每次跑的準確率都差不多，generative model則是變動很大，而且很容易變成全預測零或全預測一，由於這次的訓練資料中大部分都是零，所以generative model表現比較差可能是較容易受到random不同的影響。

| 準確率 | logistic regression | generative model |
|---------|---------------------|------------------|
| public | 0.85712 | 0.72149 |
| private | 0.85345 | 0.71907 |

2. (1%) 請說明你實作的best model，其訓練方式和準確率為何？

特徵部分，在助教處理好的123個特徵中，只有五個特徵是非0,1的，所以分別再加上這五個特徵的三次方和平方根為新的特徵，然後使用sklearn中的f_classif對這133個特徵進行篩選，經過嘗試，120個特徵時有較好的結果，最後使用LogisticRegression分類器，iter為100，正規化係數為1，有做特徵標準化(平均為零，標準差為一)時作出最好的結果，在public/private set上準確率為0.86105/0.85800。

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>)

除了不做特徵標準化以外，其他部分都和best model一致，在public/private set上結果如下表，沒有做標準化的準確率降了將近一成，由於特徵只有少數是非0,1的，且數值可以大到上萬，新加上特徵還有三次方項，和其他0,1特徵的波動比起來就會差異非常大，如果不做特徵標準化就會影響到學習效果。

| 準確率 | 有做標準化 | 沒有做標準化 |
|---------|---------|---------|
| public | 0.86105 | 0.77371 |
| private | 0.85800 | 0.77140 |

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)

除了正規化係數值以外，其他部分都和best model一致，在public/private set上結果如下表，1的時候有最好結果，但其實每個結果的差距非常小，除非把係數再加大，準確率才會明顯下降，可見對於這次的預測，正規化的影響並不大，經過篩選後的特徵都有其重要性，也不太有對訓練資料overfitting的情形。

| 係數 | 10 | 1.0 | 0.5 | 0 |
|---------|---------|---------|---------|---------|
| public | 0.85761 | 0.86105 | 0.86142 | 0.86068 |
| private | 0.85652 | 0.85800 | 0.85800 | 0.85751 |

5. (1%) 請討論你認為哪個attribute對結果影響最大？

在助教處理好的123個特徵中，只有五個特徵是非0,1的，經過實驗，加上這五個特徵的不同次方項時，準確率會有不等效果的提高，可見這五個特徵應該是相對比較重要的，因此這次實驗只針對這五個特徵，嘗試分別去掉這五個特徵的其中之一，即少了該特徵的一次方項、三次方項、平方根項，只用其餘的130個特徵來篩選出120個特徵，其他部分都和best model相同，在public/private set上結果如下表，由表格中可知，capital_gain應該是影響最大的特徵，因為去掉之後的準確率降低最多。

| 特徵 | age | fnlwgt | capital_gain | capital_loss | hours_per_week |
|---------|---------|---------|--------------|--------------|----------------|
| private | 0.85345 | 0.85763 | 0.84301 | 0.85480 | 0.85689 |
| public | 0.85761 | 0.85945 | 0.84361 | 0.85737 | 0.85737 |