學號：R07943107 系級：電子所碩一 姓名：徐晨皓

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

|  |  |  |
| --- | --- | --- |
|  | Private score | Public score |
| Generative model | 0.84080 | 0.84643 |
| Logistic regression | 0.85038 | 0.85196 |

在實作中，此兩種方法都有先對每個feature做normalization。以public score和private score而言，logistic regression表現較佳。

1. **請說明你實作的best model，其訓練方式和準確率為何？**

Best model使用的是sklearn的RandomForestClassifier(n\_estimators=500, max\_depth=20)，並且對於所有的feature都有做normalization。Public score為**0.86486**；private score為**0.86353**。

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。**

|  |  |  |
| --- | --- | --- |
|  | Private score | Public score |
| w/ normalization | 0.85038 | 0.85196 |
| w/o normalization | 0.77877 | 0.78108 |

實驗設定為logistic regression, iterations=1000, learning rate=0.5, lambda=0.001。

若沒有使用feature normalization，模型的準確率會大幅下降。原因可能是one-hot的feature為0和1，但其他實數的features (如age, fnlwgt等)是遠大於1的。這樣會使得某些features dominate整個模型。

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**。

|  |  |  |
| --- | --- | --- |
|  | Private score | Public score |
| 0 | 0.85063 | 0.85110 |
| 0.00001 | 0.85075 | 0.85110 |
| 0.0001 | 0.85050 | 0.85098 |
| 0.001 | 0.85038 | 0.85196 |
| 0.01 | 0.85014 | 0.85085 |
| 0.1 | 0.84043 | 0.84004 |
| 1 | 0.76771 | 0.77272 |
| 10 | 0.76047 | 0.76707 |
| 100 | 0.76047 | 0.76707 |

從實驗中，我們可以發現當λ小於等於0.1的時候，對準確率沒有太大的影響。但當λ大於等於1的時候，準確率就大幅下降。

1. **請討論你認為哪個attribute對結果影響最大？**

在訓練後的logistic regression模型的weights中，capital\_gain的係數(1.958)最大，而Never-married的係數(-0.4198)最小。因此，我認為capital\_gain與Never-married對結果影響最大。這也還算合理，資本收入(capital\_gain)本來就應該與個人的財產收入有很大的關係，而有無婚姻(Never-married)可能某種程度上表示個人在社會上的成熟度。