**Homework 2 Report**

系級:資工碩二 學號:R06922134 姓名:葉沛陽



|  |  |  |
| --- | --- | --- |
|  | public | private |
| Logistic regression | 0.67580 |  |
| Generative model | 0.21940 |  |

此兩個model比較基準為資料未經過one-hot encoding 且 normalization等等錢處理，直接將原始值進行訓練的結果

由表格可看出 Logistic regression 比 generative model 有比較高的準確率，可能是因為 generative model 是在假設資料是在高斯分布的前提下去運算的，再加上資料未經過one-hot encoding且normalization所以資料的分佈可能跟高斯分布差異很大。而logistic regression 則是直接使用regression方式去切割兩個class，因此至少會有一定的準確率。



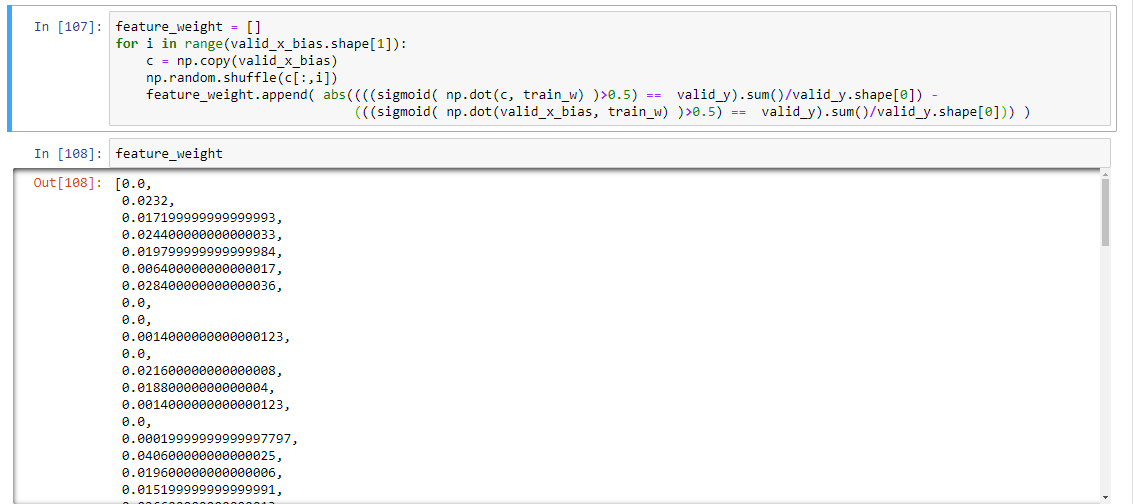
|  |  |  |
| --- | --- | --- |
|  | public | private |
| Without one-hot | 0.67580 |  |
| With one-hot | 0.78060 |  |

此表格為用同樣參數的logistic regression model，除了有無使用one-hot encoding，其他資料前處理都一致所得出的結果。

由表格可以看出有使用one-hot encoding會比沒有使用one-hot encoding來得準確，原因是因為原始資料將性別:男設定為1、女設定為2，教育…等等feature也是直接使用1234數字來表示類別，這會造成某些類別比較接近或是數值比較大，但也許跟他們事實上的關係並不符合，因此，使用one-hot encoding會讓每個類別各自獨立出來當成各個feature，這樣會比較符合事實，也可得出比較高的準確率。



我採用permutation test。簡單來說，要看某個feature影響力大不大，可以將那個feature整行隨機排列，然後觀察隨機排列前後對於結果的影響，取絕對值後，越大的表示這個feature影響結果越多越重要。



例如這樣

值為0.0表示有無隨機排列對於結果毫無影響，很不重要

而值越大表示那個feature對結果影響越大。



|  |  |  |
| --- | --- | --- |
|  | public | private |
| Without normalization | 0.78060 |  |
| With normalization | 0.81880 |  |

此表格為用同樣參數的logistic regression model，除了有無使用mean normalization，其他資料前處理都一致所得出的結果。

由表格可以看出有使用normalization會比沒有使用normalization來得準確，

原因應該是原本的資料各個feature的上下限差異非常大，也許有些feature值就算差一點點就很重要，有些feature值差很多還是沒有什麼影響，若沒有先做normalization這會造成在用梯度下降訓練weights的時候，weights會比較難match到較正確的位置，造成各feature的weight收斂速度不一，所以比較可能會得出較差的結果。

