**Machine Learning Fall 2020 ——— Homework 2**

學號：B07902037 系級： 資工三 姓名：蔡沛勳

**1. (0.5%) 請比較你實作的generative model、logistic regression 的準確率，何者較佳?**

在不對 X\_train 做任何 preprocess 的前提下，得到的結果下：

|  |  |  |  |
| --- | --- | --- | --- |
| **Training model** | **Train data score** | **Public score** | **Private score** |
| Generative model | 0.79945 | 00.80577 | 0.79523 |
| Logistic model | 0.79386 | 0.79570 | 0.79425 |

可看出兩者的表現皆不甚理想。而且不符合預期中 Logistic model 表現應優於 Generative model，反而在三項表現皆不如。可能原因為在不做 normalization 的前提下部分 feature 會嚴重影響 logistic model 的表現，導致其他 feature 無法反映出代表性。

在 normalize X\_train 後得到的結果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| Training model | Train data score | Public score | Private score |
| Generative model | 0.80421 | 0.80589 | 0.79388 |
| Logistic model | 0.85314 | 0.85393 | 0.84854 |

可看出 Generative model 變化不大的同時， Logistic model 的表現有了顯著的改善，並且也符合預測結果 —— Logistic model 表現優於 Generative model。

**2. (0.5%) 請實作特徵標準化(feature normalization)並討論其對於你的模型準確率的影響**

Normalize 方式：feature = ( feature – feature.min ) / (feature.max – feature.min)

Generative model：

|  |  |  |  |
| --- | --- | --- | --- |
| Normalize | Train data score | Public score | Private score |
| Not normalize | 0.79945 | 00.80577 | 0.79523 |
| Normalize | 0.80421 | 0.80589 | 0.79388 |

Logistic model：

|  |  |  |  |
| --- | --- | --- | --- |
| Normalize | Train data score | Public score | Private score |
| Not normalize | 0.79386 | 0.79570 | 0.79425 |
| Normalize | 0.85314 | 0.85393 | 0.84854 |

Best model (Graident boost classifier)：

|  |  |  |  |
| --- | --- | --- | --- |
| Normalize | Train data score | Public score | Private score |
| Not normalize | 0.89079 | 0.87923 | 0.87298 |
| Normalize | 0.88440 | 0.87850 | 0.87249 |

可以看出在 Generative model 及 Best model 經過 normalize 後的表現變化較不明顯，但在 Logistic model 上表現可以好到5個百分比之多。原因應為 normalize 使training Logistic model 時各項 feature 能平均的展示其對預測結果的影響，而不會使某些數字較大的 feature 影響比重大於其他數字較小的 feature。

**3. (1%) 請說明你實作的best model，其訓練方式和準確率為何?**

我用 sklearn 裡的 gradientboostclassifier 來實作 best model，參數設定如下：

|  |  |
| --- | --- |
| loss | deviance |
| learning\_rate | 0.05 |
| n\_estimators | 1000 |
| max\_depth | 5 |
| max\_features | auto |
| validation\_fraction | 0.01 |
| n\_iter\_no\_change | 10 |
| tol | 1e-4 |
| verbose | 1 |

得到結果如下：

|  |  |  |
| --- | --- | --- |
| Train data score | Public score | Private score |
| 0.88486 | 0.87886 | 0.87262 |

**4. (3%) Refer to math problem**

1.

設 所屬的 class 為 。則likelihood function為：

開 log 得到：

已知log後的likelihood function最大值與likelihood function到達最大值的 相同且 。為求出 likelihood function最大值，對 log後的likelihood function使用 Lagrange multiplier。設 ， 。

對 用 偏微分得

因此

對 用 偏微分得

因此

令 ，則得到

已知，可推得

可得 maximun-likelihood solution為

2.

由題目得到

3.

與第一題相同，求 log likelihood function 的最大值，得到

已知 Normal distribution 可表示為

對 用 偏微分得

極值在偏微分等於0，故得出

得證 maximun-likelihood solution 之為

對 用 偏微分得

其中間算式可透過第二題推得

後面算式為

故合併得

極值在偏微分等於0，故得出

得證 maximun-likelihood solution 之為