Machine Learning 2019 Spring - HW2 Report

學號：B06902029 系級：資工二 姓名：裴梧鈞

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

Model	Public Score	Private Score
Generative	0.84004	0.83540
Logistic Regression	0.85233	0.85136

Logistic Regression 有些微較高的正確率。

2. **請說明你實作的best model，其訓練方式和準確率為何？**

Model	Public Score	Private Score
Gradient Boosting	0.87641	0.87483

我使用的 Model 是 Gradient Boosting

1. 在處理資料時，我有做 feature normalization，並把 `fnlwgt` 這項 feature 拿掉。
2. 在連續的 feature，像是 *"age", "capital_gain", "capital_loss", "hours_per_week"*，我有加入二次及三次項
3. 我做 Gradient Boosting 的參數是
    1. `n_estimators`：173
    2. `learning_rate`：0.05
    3. `max_depth`：6
    4. `random state`：將我的名字 "Wu-Jun Pei" 做 sha256sum 轉成整數模 2^{32}

在選擇參數時，我有枚舉這些參數，並使用 `cross_val_score` 綜合選擇出最好的！

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。**

我在前三的實作（包含 Gradient Boosting、Logistic Regression、Generative Model）都是有實作 feature normalization 的。在此，我使用的是 Logistic Regression 的 model，在沒有調整任合參數（如 learning rate、optimizer 等）的情形下，比較 feature normalization 的影響。

Feature Normalization	Public Score	Private Score
True	0.85233	0.85136
False	0.85245	0.85149

可以看到 Feature Normalization 前的 Score 反而略高一些，我認為一種可能的原因是

1. 資料有相當多因為 one-hot encoding 而使用 0/1 作為 feature，可以觀察到僅有 4 種 feature 是連續的，他們的影響可能沒有很大
2. 我的 epochs 次數夠多，導致沒有 normalization 的 model 也走到一個不錯的最低點

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。**

> 在這題，我使用的 Model 是 Logistic Regression，有使用 Feature Normalization，並使用 Adam 優化，epochs 次數則設成 10000。

$\log_{10} \lambda$	**Training Accuracy**	**Public Score**	**Private Score**
-2	0.85243	0.85233	0.85136
-1	0.85243	0.85233	0.85136
0	0.85236	0.85221	0.85149
1	0.85270	0.85282	0.85136
2	0.85141	0.85380	0.85050

我認為這次的作業，regularization 並沒有很大的影響，training accuracy 和 public score、private score 都沒有相差很多。在我測試的前四個例子中，regularization的影響可說是微乎其微；在第五個例子，可以看到這個 model 可能已經 underfit 了，training accuracy 和 private score 都相對顯著下降，而 public score 的上升我認為可以視為一種運氣、隨機因素，上升的幅度也不大。

5. **請討論你認為哪個 attribute 對結果影響最大？**

> 和第四題一樣，這題我使用的 Model 是 Logistic Regression，有使用 Feature Normalization，並使用 Adam 優化，epochs 次數則設成 10000。

我有嘗試將 $\mathbf{w}$ 印出來，並試著比較各自的差距，並使用各個 feature 的 $|\mathbf{w}_i|$ 排序，前五名結果如下：

```
 2.3601675130281428, 'capital_gain'
-0.7015782027960535, ' Never-married'
-0.6903933475767674, ' Preschool'
 0.5954210921730702, ' Married-civ-spouse'
 0.4090251657216704, 'sex'
```

可以發現 `capital_gain` 佔有相當大的影響力