作業一 報告

資訊 114 H44091196 洪茂菘

練習 1：改善決策樹分類模型

這裡我分別採用了兩種方法來試著改善 test acc 預測原始分數，以下會個別對改進方法及改進部分作介紹講解:

- **作法一: 使用超參數**

如下圖所示，這邊我仿照助教上課的教學來設定超參數，我將最佳化方法設定為 Gini Index；最大設定深度為 10，最多葉子個數則設定為 5^10(以上數字是我自己實測出來的)。

```
model = DecisionTreeClassifier(
        criterion='gini',
        max_depth=10,
        max_leaf_nodes=5 ** 10,
)
```

使用後的前後結果差異如下:

```
train accuracy: 0.9353932584269663
test accuracy: 0.7821229050279329
```

```
train accuracy: 0.9831460674157303
test accuracy: 0.7262569832402235
```

如上兩張圖，第一張圖是使用超參數後的結果，第二張則是原本的範例，可以觀察到 test accuracy 從原本的 0.7262 提升至 0.7821。

- **作法二: 增加更多的輸入特徵**

```
df_x = df[['Sex', 'Age', 'Fare', 'Pclass', 'SibSp', 'Parch']]
```

在助教提供的範例中，原先只有 Sex, Age, Fare 三個輸入特徵，我自己則是新增了後三個特徵來試著提升 test acc 的預測原始分數，以下會簡單介紹我選擇該參數的原因:

1. Pclass:

   根據資料解釋，Pclass 屬性為鐵達尼號中乘客所搭乘的船艙等級，分為 1,2,3，數字越小則代表越高級，因此，我認為待在艙等越高的乘客，其生存率也會越高，實際對測試資料統計下來的結果也是如

此，故將其列入輸入特徵中。

2. SibSp 和 Parch:

   根據資料，**SibSp 代表的是「手足和配偶人數」；Parch 則代表「父母及子女人數」**，由於這兩個屬性的性質較類似，所以我放在一起解釋。

   依據我對測試資料的觀察與實際統計結果，我發現 SibSp 和 Parch 人數越少的乘客(即同行家人越少)，其生存率較高，推測其原因可能是因為當有同行家人時，在發生船難的當下第一反應是設法去拯救自己的家人，進而導致自己的存活率降低了。由以上分析，我最終也選擇將這兩個屬性列入至輸入特徵中。

使用後的前後結果差異如下:

```
train accuracy: 0.9859550561797753
test accuracy: 0.7430167597765364
```

```
train accuracy: 0.9831460674157303
test accuracy: 0.7262569832402235
```

如上兩張圖，第一張圖是使用增加輸入特徵後的結果，第二張則是原本的範例，可以觀察到 test accuracy 從原本的 0.7262 略微提升至 0.7430。

練習 2: 使用不同的模型

第二小題則是要求我們使用不同的模型來試著超越原本的 test accuracy 數值。這邊我選了兩個分數表現較原本高的模型來呈現:

1. ensemble.RandomForestClassifier 模型

```
#  創造決策樹模型
model  =  RandomForestClassifier(random_state=1012)
```

```
train accuracy: 0.9831460674157303
test accuracy: 0.7486033519553073
```

如上兩張圖所示，我第一個選用的是集合學習中的 RandomForestClassifier 模型，可以看到其結果為 0.7486，大於原本的 0.7262。

2. ensemble. ExtraTreesClassifier 模型

```
#  創造決策樹模型
model  =  ExtraTreesClassifier(random_state=1012)
```

```
train accuracy: 0.9831460674157303
test accuracy: 0.7430167597765364
```

如上兩張圖所示，我第二個選用的也是集合學習中的 ExtraTreesClassifier 模型，可以看到其結果為 0.7430，大於原本的 0.7262。