Decision Tree Report

110605006 資工三B 劉韶颺

1. Function Explanation--

_feature_split

- _feature_split 函數的目標是在給定輸入資料 X 和目標 y 的情況下，查找最佳的資料分割點。遍歷所有資料，以找到產生最小亂度的分割點，最後返回最佳分割點的特徵索引feature_idx和閾值threshold。

_build_tree

- 目標是遞歸地構建決策樹，以_feature_split選擇最佳的特徵分割點，然後分割數據，形成樹的節點。這個過程重複進行，直到達到最大深度，將決策樹完整建構。

_find_min_alpha

- 目標是在測試決策樹時，遞歸地檢查每個節點，計算其 alpha 值找到擁有最小 alpha 值的節點，以進行最佳的剪枝操作，最後返回具有最小 alpha 值的節點，以進行後續剪枝。

_prune

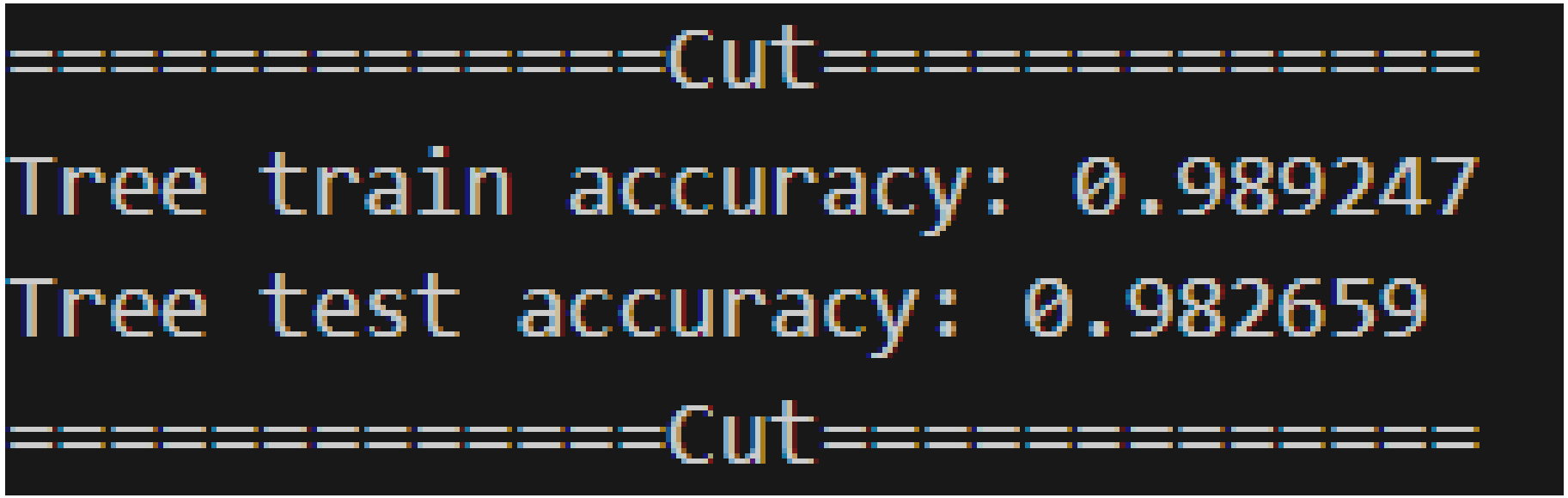
- 目標是實際執行剪枝操作，刪除從_find_min_alpha回傳之擁有最小 alpha 值的節點，將選定的節點的左右子樹設為空，實現剪枝操作以簡化樹的結構。這個過程有助於減少模型的複雜性，同時減少overfitting的風險。

2. Decision tree before post-pruning accuracy

```
[Running] python -u "c:\Users\Benson\OneDrive - 國立中央大學\桌面\資料科學導論
\code_110605006\code_110605006.py"
Tree train accuracy: 1.000000
Tree test accuracy: 0.990366
=============Cut=============
```

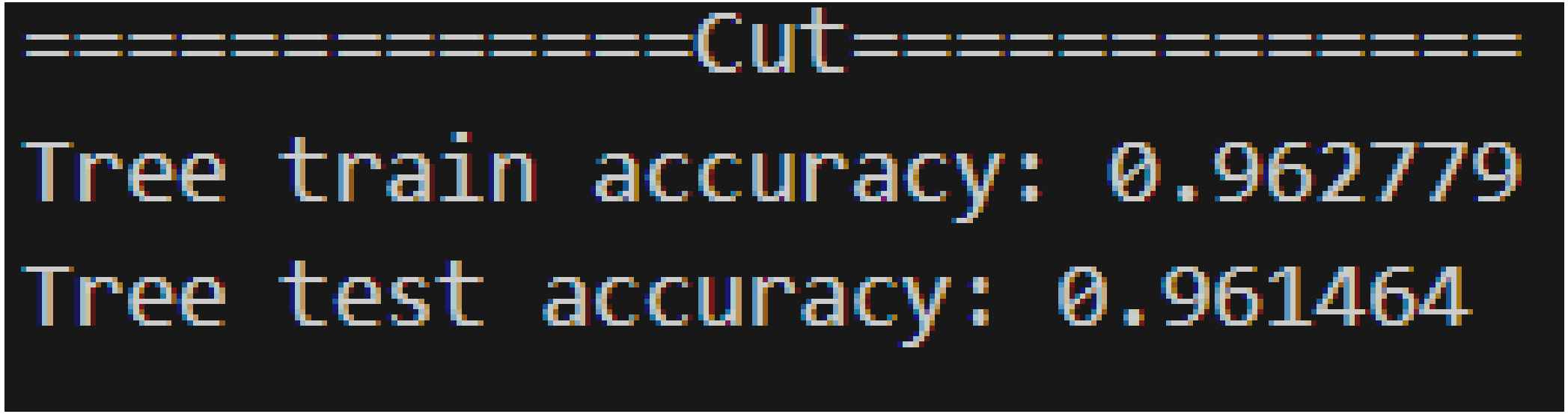
3. Decision tree after post-pruning accuracy

after 10 times:



3. Decision tree after post-pruning accuracy

after 20 times:



4. The effect of different parameters

- Prune tree times:

  prune tree times增加，準確率下降。

- Max_depth:

  經過測試發現通常max_depth越大，準確率越高，不過會有上限；以test_ratio為0.2為例，max_depth越大，準確率越高，不過當max_depth>=14後，準確率就不會再更高了。

- Test_ratio:

  較大的測試集可以提供更可靠的性能評估，但也可能減少訓練數據的量;較小的測試集可能導致overfitting的風險

5. A brief discussion of the results

- 以我的程式碼測試，不論是train或是test都會因prune tree times的增加而造成準確率下降；雖然修剪應該是為了減少overfitting的風險，不過我認為這次的結果是因為剪枝過多，導致decision tree過於簡化而逐漸失去數據劃分能力，因此，應謹慎權衡剪枝程度來達到最大效能。

- 在測試test_ratio時，我發現較大的測試集可以提供更可靠的性能評估，但也可能減少訓練數據的量;較小的測試集可能導致overfitting的風險，不過我認為在比較或測試其他參數時，使用固定的test_ratio會比較好