HW1.Decision Tree

Tree Structure

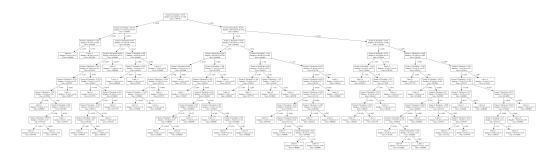


Figure 1: Tree Structure

Discussion

- 1. 圖片若模糊不清可參考資料夾中的「tree_structure.svg」
- 2. 決策樹參數:max_depth=10, min_leaves=5, 準確率為 83.6193%
- 3. Figure 2 是不同最小子葉、最大深度測試結果的折線圖,詳細表格可以參考附錄中的 Table $1 \circ$

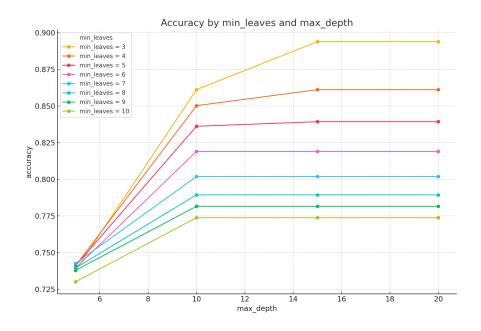


Figure 2: 訓練資料與測試資料相同

可以發現,在最小子葉數為 3,最大深度為 20 時,準確率有最高的 89.39%,準確率極高,但這樣的模型可能會過擬合。

於是我又進行了一個測試:隨機取用 620 筆資料訓練,20 筆資料測試且完全不重複(測試資料儲存在 testdatas 資料夾,使用 gen.cpp 生成),在每組不同最小子葉數、最大深度的組合都進行 20 次測試,取其平均值以求精確。結果如 Table 2,結果折線圖如 Figutre3:

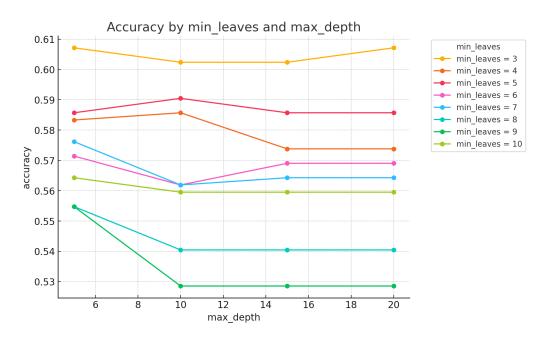


Figure 3: 訓練資料與測試資料完全不同

最高的準確率為 60.71%,和原本的結果差距極大,顯然是過擬合的問題。

How to improve accuracy

解決過擬合的方法有很多,例如增加訓練資料、減少特徵數、增加正則化等。但在測試資料有限的狀況下,我選擇使用隨機森林來解決過擬合的問題:建構多棵決策樹,並將每棵決策樹的結果進行投票,最後選擇投票最多的結果作為最終結果。

以下是隨機森林測試的結果,表格資料可以參考附錄中的 Table 3:

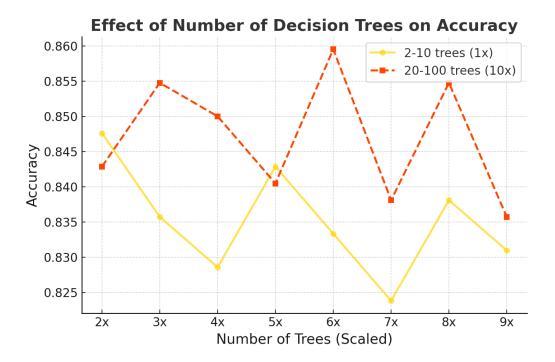


Figure 4: 隨機森林測試結果

我建構 $2\sim10$, $20\sim100$ 棵決策樹,每棵決策樹的最大深度為 10,最小子葉數為 5 進行測試。黃色實線代表 $2\sim10$ 棵決策樹的準確率;橘色虛線則是 $20\sim100$ 棵決策樹的準確率,以上的測試均使用 20 筆不同的測試資料取其平均。PS. 由於隨機森林的取樣存在隨機性,因此我原本對於每組測試資料要建構 10 次隨機森林取平均,但這樣下來總共需要建構 118800 棵決策樹,我的筆電無法負荷這麼大的運算量(測試時執行了一個多小時才完成五分之一),因此我只建構了一次

可以發現,即使是最差的狀況也有八成以上的準確率,能有效解決過擬合的問題。

References

1. https://medium.com/@SCU.Datascientist/python 學習筆記-決策樹-decision-tree-b9acf11f0f84

- $2. \ https://ithelp.ithome.com.tw/articles/10271143?sc=hot$
- 3. https://zh.wikipedia.org/zh-tw/決策樹學習
- $4.\ https://blog.csdn.net/qq_38502736/article/details/107210625$
- 5. https://github.com/mcxiaoxiao/c-Decision-tree
- 6. https://zh.wikipedia.org/zh-tw/隨機森林
- 7. https://ithelp.ithome.com.tw/m/articles/10272586
- 8. ChatGpt (協助建立折線圖、表格、letex 排版)

Appendix

Tables of Decision Trees

Table 1: 使用決策樹,訓練資料與測試資料完全相同

最小子葉	最大深度	準確率	最小子葉	最大深度	準確率
3	5	73.79 %	7	5	74.26 %
3	10	86.12~%	7	10	80.19~%
3	15	89.39~%	7	15	80.19~%
3	20	89.39~%	7	20	80.19~%
4	5	74.10~%	8	5	73.95~%
4	10	85.02~%	8	10	78.94~%
4	15	86.12~%	8	15	78.94~%
4	20	86.12~%	8	20	78.94~%
5	5	74.10~%	9	5	73.79 %
5	10	83.62~%	9	10	78.16~%
5	15	83.93~%	9	15	78.16~%
5	20	83.93~%	9	20	78.16~%
6	5	73.95~%	10	5	73.01~%
6	10	81.90~%	10	10	77.38~%
6	15	81.90~%	10	15	77.38~%
6	20	81.90 %	10	20	77.38 %

Table 2: 使用決策樹,訓練資料與測試資料完全不同

最小子葉	最大深度	準確率	最小子葉	最大深度	準確率
3	5	60.71 %	7	5	57.62 %
3	10	60.24~%	7	10	56.19~%
3	15	60.24~%	7	15	56.43~%
3	20	60.71~%	7	20	56.43~%
4	5	58.33~%	8	5	55.48~%
4	10	58.57~%	8	10	54.05~%
4	15	57.38~%	8	15	54.05~%
4	20	57.38~%	8	20	54.05~%
5	5	58.57~%	9	5	55.48~%
5	10	59.05~%	9	10	52.86~%
5	15	58.57~%	9	15	52.86~%
5	20	58.57~%	9	20	52.86~%
6	5	57.14~%	10	5	56.43~%
6	10	56.19~%	10	10	55.95~%
6	15	56.90~%	10	15	55.95~%
6	20	56.90 %	10	20	55.95 %

Tables of Random Forests

Table 3: 使用隨機森林,訓練資料與測試資料完全不同

決策樹數量	準確率	決策樹數量	準確率
2	0.847619	10	0.852381
3	0.835714	20	0.842857
4	0.828571	30	0.854762
5	0.842857	40	0.850000
6	0.833333	50	0.840476
7	0.823810	60	0.859524
8	0.838095	70	0.838095
9	0.830952	80	0.854762
90	0.835714	100	0.838095