機器學習觀念與應用 作業三

112703003 資訊二 黃柏淵

第一題：給定 training_data.csv，請先針對資料進行資料前處理，參考課堂範例，經由 One-Hot Encoding 後，除了標籤欄位，(1)用於訓練 Decision Tree 模型的欄位共有幾個？(2)請依序列出各欄位名稱與該欄位說明（依照欄位名稱字母順序升冪排序）。

(1)15 個

(2)如下表

buying_low	車輛購買價格為 low
buying_med	車輛購買價格為 med
buying_vhigh	車輛購買價格為 vhigh
doors_3	車門數量為 3
doors_4	車門數量為 4
doors_5more	車門數量為五個或以上
lug_boot_med	行李箱大小為 med
lug_boot_small	行李箱大小為 small
maint_low	維護費用為 low
maint_med	維護費用為 med
main_vhigh	維護費用為 vhigh
persons_4	可搭乘人數為 4 人
persons_more	可搭乘人數為 more(超過 5 人)
safety_low	安全性為 low
safety_med	安全性為 med

第二題：承 1，本題任務使用 Entropy 作為 Impurity Metric，請在不使用 Early-Stop Rules 的情況下，使用全部 300 筆資料生成 The Fully-Grown Decision Tree（即不限制 Decision Tree 的 Max. Depth、Max. Number of Leaf Nodes、Min.Number of Instances 等），請列出此 Decision Tree 的 (1) Max Depth 和(2)Leaf Nodes 總數。[注意：使用 DecisionTreeClassifier 時，僅設定 criterion='entropy'，其餘使用 DecisionTreeClassifier 的預設參數。]

```
Max Depth: 10
Leaf Nodes: 37
```

第三題：承 1，為了有效建立分類模型，以及評估模型分類的效果，我們採用 Holdout 策略，練習使用 sklearn.model_selection 的 train_test_split 將已有的 300 筆資料分成 70% 為訓練集和 30% 為測試集，再進行模型訓練，使用 train_test_split 時，僅指定 test_size=0.3、random_state=42，其餘使用 train_test_split 的預設參數，本題任務請使用 Entropy 作為 Impurity Metric，在不使用 Early-Stop Rules 的情況下，使用訓練集 210 筆資料生成 The Fully-Grown Decision Tree，請列出此 Decision Tree 的 (1) Max Depth、(2) Leaf Nodes 總數、(3) 所有的 Internal Nodes 的 Index、Attribute/Feature Name、Split Threshold (輸出請依照 Node Index 升冪排序，參考圖一)。

(1、2)

```
Max Depth: 9
Leaf Nodes: 27
```

(3)

```
Internal Node Index: 0
  feature_name: safety_low
  split threshold: 0.5000
----------------------------
Internal Node Index: 1
  feature_name: buying_vhigh
  split threshold: 0.5000
----------------------------
Internal Node Index: 2
  feature_name: maint_vhigh
  split threshold: 0.5000
----------------------------
Internal Node Index: 3
  feature_name: persons_4
  split threshold: 0.5000
----------------------------
Internal Node Index: 4
  feature_name: persons_more
  split threshold: 0.5000
----------------------------
Internal Node Index: 6
  feature_name: lug_boot_small
  split threshold: 0.5000
----------------------------
Internal Node Index: 8
  feature_name: doors_3
  split threshold: 0.5000
----------------------------
```

```
Internal Node Index: 9
  feature_name: doors_5more
  split threshold: 0.5000
----------------------------
Internal Node Index: 10
  feature_name: doors_4
  split threshold: 0.5000
----------------------------
Internal Node Index: 15
  feature_name: safety_med
  split threshold: 0.5000
----------------------------
Internal Node Index: 17
  feature_name: lug_boot_small
  split threshold: 0.5000
----------------------------
Internal Node Index: 18
  feature_name: lug_boot_med
  split threshold: 0.5000
----------------------------
Internal Node Index: 20
  feature_name: doors_4
  split threshold: 0.5000
----------------------------
Internal Node Index: 21
  feature_name: buying_low
  split threshold: 0.5000
----------------------------
```

```
Internal Node Index: 25
  feature_name: maint_low
  split threshold: 0.5000
----------------------------
Internal Node Index: 26
  feature_name: doors_3
  split threshold: 0.5000
----------------------------
Internal Node Index: 30
  feature_name: doors_3
  split threshold: 0.5000
----------------------------
Internal Node Index: 31
  feature_name: buying_low
  split threshold: 0.5000
----------------------------
Internal Node Index: 33
  feature_name: persons_more
  split threshold: 0.5000
----------------------------
Internal Node Index: 35
  feature_name: lug_boot_med
  split threshold: 0.5000
----------------------------
Internal Node Index: 38
  feature_name: lug_boot_med
  split threshold: 0.5000
----------------------------
```

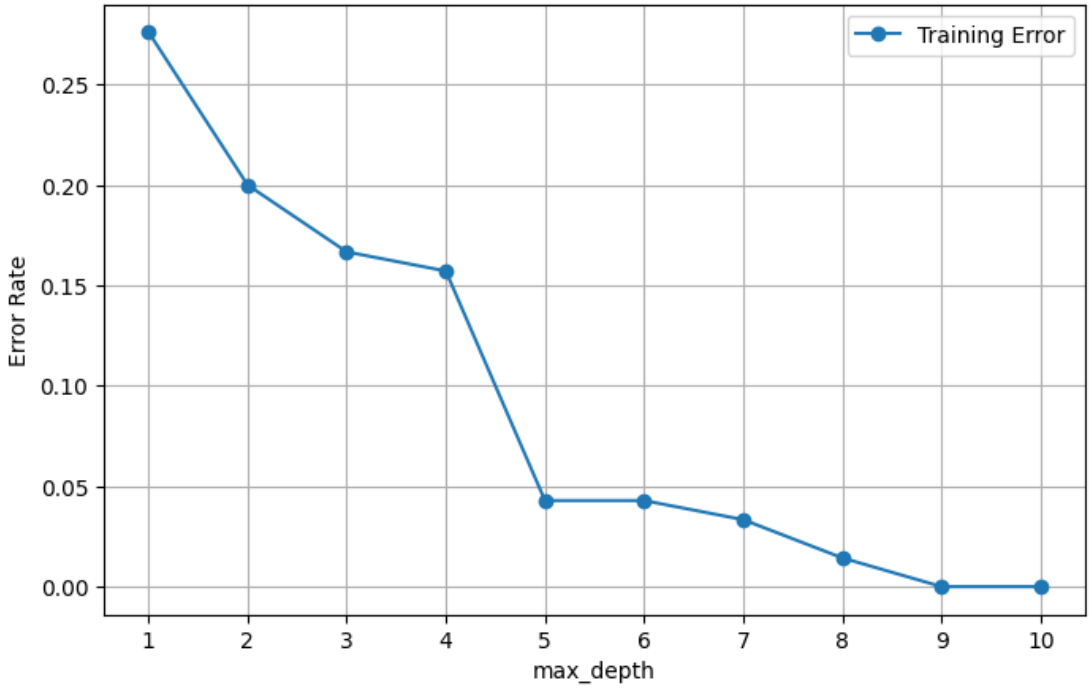
```
Internal Node Index: 39
  feature_name: persons_4
  split threshold: 0.5000
----------------------------
Internal Node Index: 40
  feature_name: lug_boot_small
  split threshold: 0.5000
----------------------------
Internal Node Index: 45
  feature_name: maint_med
  split threshold: 0.5000
----------------------------
Internal Node Index: 47
  feature_name: persons_more
  split threshold: 0.5000
----------------------------
Internal Node Index: 49
  feature_name: lug_boot_small
  split threshold: 0.5000
----------------------------
```

第四題：承 3，利用已建立的 Decision Tree 模型，計算 (1) Training Error 為？(即訓練集 210 筆資料的錯誤率) (2) Test Error 為？(即測試集 90 筆資料的錯誤率)

```
Training Error: 0.0000
Test Error: 0.0556
```

第五題：承 4，本題任務練習使用 Early-Stop Rules，使用 DecisionTreeClassifier 時，僅指定 criterion='entropy' 和 max_depth 數值為 1~10 的整數，其餘使用 DecisionTreeClassifier 的預設參數，並觀察 Training Error 的變化。(1)請提供 max_depth 和 Training Error 的關係曲線圖，如圖二，(2) 說明 max_depth 和 Training Error 的關係變化，並解釋為何有此現象。

(1)如下圖：



(2)趨勢：隨著 max_depth 從 1 增加到 10，Training Error 整體呈現下降趨勢。在 max_depth 較小時 (1~4)，錯誤率下降較快。當 max_depth 達到 5 時，錯誤率已大幅降低，並在 5 到 10 之間趨於穩定，下降幅度變小。

原因：max_depth = 1 時，整棵樹只有一個分裂點。模型只能根據一個條件，把整個資料集粗略地劃分成兩群。大部分資料會被分錯，因為只靠一個條件無法充分描述資料的複雜性，所以錯誤率很高。增加 max_depth，Decision Tree 可以做更多次的細分。每多一層，資料被更細緻地切分（每個子集更乾淨）。每一次切分，Training Error 都會下降，因為模型可以針對資料的細節做更好的分類。當 max_depth 足夠大時，Decision Tree 可以一直切到只剩下單一類型的資料。Training Error 趨近 0。

第六題：承 5，請利用 Nested Cross-Validation，觀察不同 max_depth 值和 Errorval 的變化，當 max_depth 的數值為 1~10 的整數時，請試著從中挑選 max_depth 值應該設為多少？

```
Overall Outer CV Accuracy: 0.9133
建議的 max_depth (取各外部折中出現最多次者): 7
```

第七題：承 6，請針對所選擇的 max_depth，使用訓練集 210 筆資料，生成 Decision Tree，在使用 DecisionTreeClassifier 時，僅指定 criterion='entropy' 和 max_depth 數值，請列出此 Decision Tree 的 (1) Leaf Nodes 總數，(2) Training Error(即訓練集 210 筆資料的錯誤率)，(3)Test Error(即測試集 90 筆資料的錯誤率)。

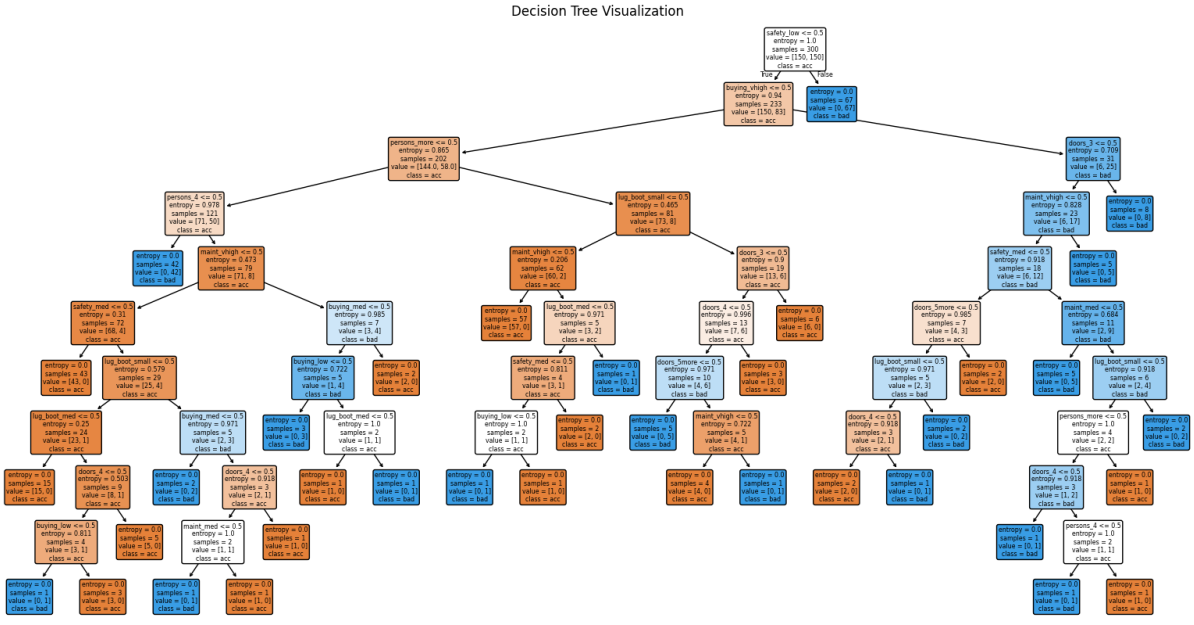
```
Leaf Nodes: 22
Training Error: 0.0333
Test Error: 0.0889
```

第八題：利用給定的 training_data.csv，自行訓練 Decision Tree 的二元分類模型，並提供此 Decision Tree 的 (1) Max Depth、(2) Leaf Nodes 總數、(3) Decision Tree 視覺化圖，如圖三。請利用此 Decision Tree 模型，針對 P3_test.csv 的測試資料，依序預測每一筆的標籤(bad/acc)，並產生 submission.csv，且將此檔案提交至 Moodle，輸出格式請參考 submission_template.csv。

(1、2)

```
Max Depth: 10
Leaf Nodes: 37
```

(3)



註明：部分程式碼及圖像使用 Grok 及 ChatGPT 輔助