Hw 3 Report

Hw 3 - Decision Tree

AU7008 Data Mining, SJTU, 2023 Spring

By **Prof. X. He**

Table of Contents

Problem Description

For the following data,

ID	A_1	A_2	A_3	*Class*
1	T	T	1	*P*
2	T	T	6	*P*
3	T	F	5	*N*
4	F	F	4	*P*
5	F	T	7	*N*
6	F	T	3	*N*
7	F	F	8	*N*
8	T	F	7	*P*
9	F	T	5	*N*

1. Train a binary decision tree;
2. using the **Gini index** metric to judge, which of A_1 and A_2 is the best split;
3. calculate the **entropy** metric to determine the best split of the continuous attribute A_3;
4. using the **entropy** metric to judge, which of A_1, A_2 and A_3 is the best split.

Solution

Question-1

By applying *Hunt's Algorithm*, we have the following binary decision tree,

```
A1?
 |-------------------------
 |                        |
 T                        F
 |                        |
A2?                      A2?
 |----------              |----------
 |         |              |         |
 T         F              T         F
 |         |              |         |
 P        A3?             N        A3?
           |-----                   |-----
           |    |                   |    |
           =5   =7                  =4   =8
           |    |                   |    |
           N    P                   P    N
```

Question-2

There are 4 occurrence as *P* and another 5 as *N*, and thus yielding a root Gini index of $1 - (4/9)^2 - (5/9)^2 = 40/81$.

Occurrence for A_1 and A_2 after splitting are,

	$N_1 : A_1 = T$	$N_2 : A_1 = F$	$N_3 : A_2 = T$	$N_4 : A_2 = F$
P	3	1	2	2
N	1	4	3	2

yielding,

- $I(N_1) = 1 - (3/4)^2 - (1/4)^2 = 3/8$
- $I(N_2) = 1 - (1/5)^2 - (4/5)^2 = 8/25$
- $I(N_3) = 1 - (2/5)^2 - (3/5)^2 = 12/25$
- $I(N_4) = 1 - (2/4)^2 - (2/4)^2 = 1/2$

further yielding,

- $\Delta_{A_1} = 40/81 - \frac{4}{9}I(N_1) - \frac{5}{9}I(N_2) = 40/81 - \frac{4}{9} \times 3/8 - \frac{5}{9} \times 8/25 = 121/810 \approx 0.149$
- $\Delta_{A_2} = 40/81 - \frac{5}{9}I(N_3) - \frac{4}{9}I(N_4) = 40/81 - \frac{5}{9} \times 12/25 - \frac{4}{9} \times 1/2 \approx 4.94 \times 10^{-3}$

Since $\Delta_{A_1} > \Delta_{A_2}$, A_1 is the best split.

Question-3

For A_3.

	≤ 0	> 0	≤ 1	> 1	≤ 3	> 3	≤ 4	> 4	≤ 5	> 5	≤ 6	> 6	≤ 7	> 7	≤ 8	> 8
P	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
N	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0

yielding entropy values for splitting by each partition value,

- 0: $E_0 = \frac{9}{9} \times [-\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9}] \approx 0.298$
- 1: $E_1 = \frac{1}{9} \times [-\frac{1}{1}\log\frac{1}{1}] + \frac{8}{9} \times [-\frac{3}{8}\log\frac{3}{8} - \frac{5}{8}\log\frac{5}{8}] \approx 0.255$
- 3: $E_3 = \frac{2}{9} \times [-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}] + \frac{7}{9} \times [-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}] \approx 0.298$
- 4: $E_4 = \frac{3}{9} \times [-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}] + \frac{6}{9} \times [-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}] \approx 0.276$
- 5: $E_5 = \frac{5}{9} \times [-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}] + \frac{4}{9} \times [-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}] \approx 0.296$
- 6: $E_6 = \frac{6}{9} \times [-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}] + \frac{3}{9} \times [-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}] \approx 0.293$
- 7: $E_7 = \frac{8}{9} \times [-\frac{4}{8}\log\frac{4}{8} - \frac{4}{8}\log\frac{4}{8}] + \frac{1}{9} \times [-\frac{1}{1}\log\frac{1}{1}] \approx 0.268$
- 8: $E_8 = \frac{9}{9} \times [-\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9}] \approx 0.298$

By choosing the minimum to maximize the split difference, **1 is the best split**.

Question-4

Occurrence for A_1 and A_2 after splitting are,

	$N_1 : A_1 = T$	$N_2 : A_1 = F$	$N_3 : A_2 = T$	$N_4 : A_2 = F$
P	3	1	2	2
N	1	4	3	2

yielding entropy values,

- $E(N_1) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.2442$
- $E(N_2) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} \approx 0.2173$
- $E(N_3) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \approx 0.2923$
- $E(N_4) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} \approx 0.3010$

further yielding,

- $E_{A_1} = \frac{4}{9}E(N_1) + \frac{5}{9}E(N_2) \approx 0.229$
- $E_{A_2} = \frac{5}{9}E(N_3) - \frac{4}{9}E(N_4) \approx 0.296$

By choosing the minimum of,

- $E_{A_1} = \frac{4}{9}E(N_1) + \frac{5}{9}E(N_2) \approx 0.229$
- $E_{A_2} = \frac{5}{9}E(N_3) - \frac{4}{9}E(N_4) \approx 0.296$
- $E_{A_3} = E_1 \approx 0.255$

to maximize the split difference, A_1 **is the best split**.