

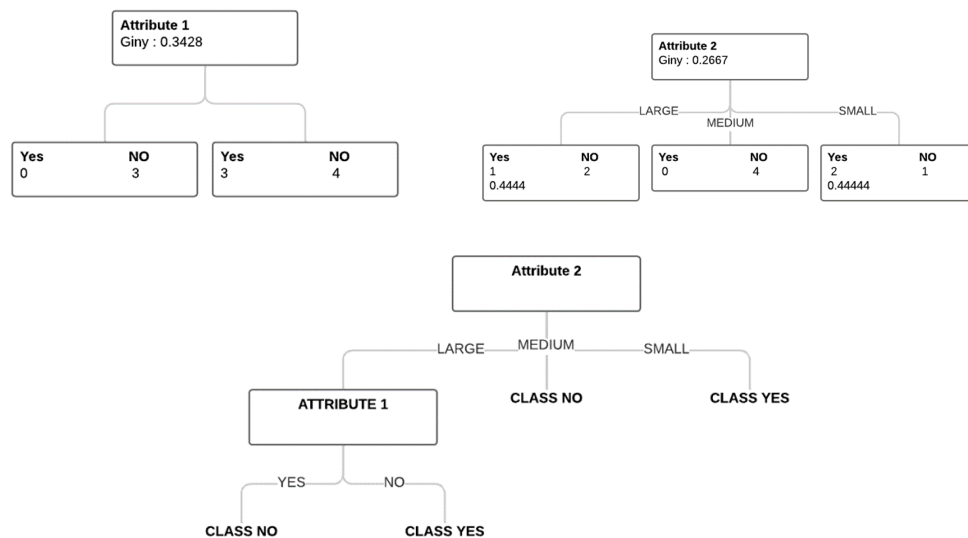
Machine Learning HW 10

DECISION TREES

Bui Phuong Thao - 11215341 - DSEB 63

1 Exercise 1

Build a decision tree based on given dataset using GINI IMPURITY



ATTRIBUTE 1	ATTRIBUTE 2	CLASS
NO	SMALL	YES
YES	MEDIUM	NO
YES	LARGE	NO
NO	SMALL	YES
NO	LARGE	YES

Build a decision tree based on given dataset using ENTROPY

$$\text{Entropy}(\text{Class}) = E(7,3) = -\left(\frac{7}{10} \log_2 \frac{7}{10}\right) - \left(\frac{3}{10} \log_2 \frac{3}{10}\right) = 0.881$$

$$\text{Entropy}(\text{Class}, \text{Attribute 1}) = P(\text{Yes}) \cdot E(\text{Yes}) + P(\text{No}) \cdot E(\text{No}) = \frac{3}{10} \cdot E(0, 3) + \frac{7}{10} \cdot E(4, 3) = 0 + 0.6896 = 0.6896$$

$$\begin{aligned} \text{Entropy}(\text{Class}, \text{Attribute 2}) &= P(\text{Large}) \cdot E(\text{Large}) + P(\text{Medium}) \cdot E(\text{Medium}) \\ &+ P(\text{Small}) \cdot E(\text{Small}) = \frac{3}{10} \cdot E(2, 1) + \frac{4}{10} \cdot E(4, 0) + \frac{3}{10} \cdot E(1, 2) = 0.550 \end{aligned}$$

Information gain for each split:

$$\text{Gain}(\text{Class}, \text{Attribute 1}) = E(\text{Class}) - \text{Entropy}(\text{Class}, \text{Attribute 1}) = 0.881 - 0.689 = 0.192$$

$$\text{Gain}(\text{Class}, \text{Attribute 2}) = E(\text{Class}) - \text{Entropy}(\text{Class}, \text{Attribute 2}) = 0.881 - 0.550 = 0.331$$

→ Choose Attribute 2 for the first split

2 Exercise 2

Humidity	Play Tennis?	
59	No	59
68	Yes	68
72	Yes	72
74	Yes	74
77	Yes	77
79	Yes	79
80	Yes	80
87	No	87
89	Yes	89
90	No	90
91	No	91
93	Yes	93
96	Yes	96
97	No	97

59	$63.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.1131$
68	$70 \rightarrow 0.94 - \frac{2}{10} \cdot 1 - \frac{12}{14} \cdot \left(-\frac{8}{14} \cdot \log_2 \frac{8}{14} - \frac{4}{14} \cdot \log_2 \frac{4}{14}\right) = 0.01$
72	$73 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.0004$
74	$75.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.015$
77	$78 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.045$
79	$79.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.09$
80	$83.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.152$
87	$88 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.048$
89	$89.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.102$
90	$90.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.025$
91	$92 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.0004$
93	$94.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.01$
96	$96.5 \rightarrow 0.94 - 0 \cdot \frac{9}{13} \cdot \log_2 \frac{9}{13} - \frac{4}{13} \cdot \log_2 \frac{4}{13} = 0.113$
97	

Handling numerical attributes

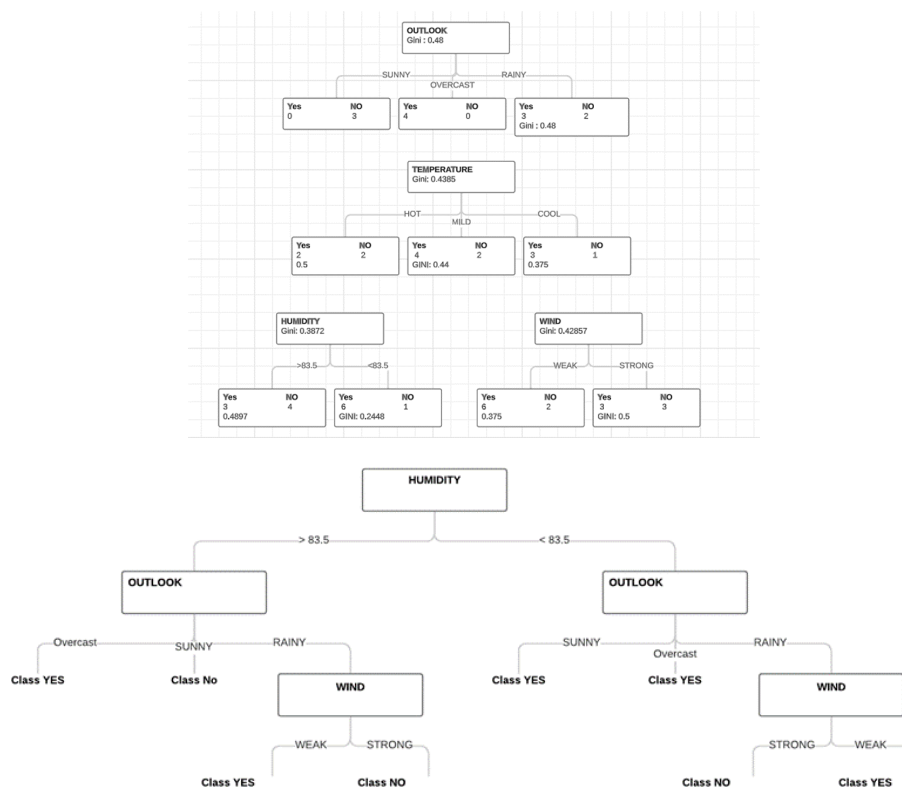
* sort the data in ascending order

* calculate the average of each adjacent pair of values

* compute the information gain with each splitting value to find the largest one

83.5 is the best splitting value with an information gain of 0.152, Humidity is now treated as a categorical attribute with two possible values

3 Exercise 3



Sample (Sunny, mild, 85, weak) belongs to class YES