Nguyen Vu Anh Ngoc
Machine Learning 1
November 23, 2023

Homework Week 10: Decision Tree

Question 1. Building the Decision Tree

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$\Delta\text{Gini} = \text{Gini}(D) - \text{Weighted Gini}(D)$$

$$\text{Weighted Gini}(D) = \sum_{j=1}^{k} \left(\frac{|D_j|}{|D|}\right) \times \text{Gini}(D_j)$$

$$\text{Entropy}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

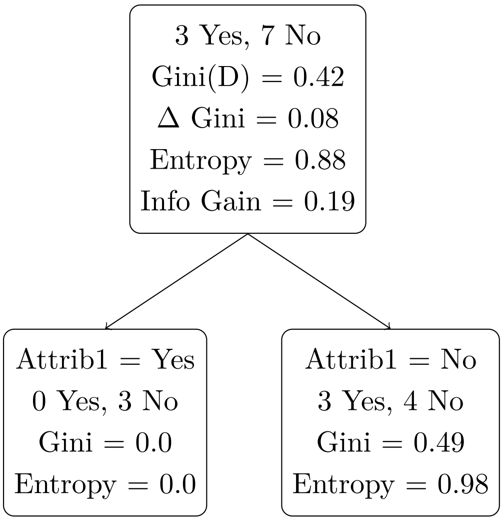
$$\text{Information Gain} = \text{Entropy}(D) - \text{Weighted Entropy}(D)$$

$$\text{Weighted Entropy}(D) = \sum_{j=1}^{k} \left(\frac{|D_j|}{|D|}\right) \times \text{Entropy}(D_j)$$

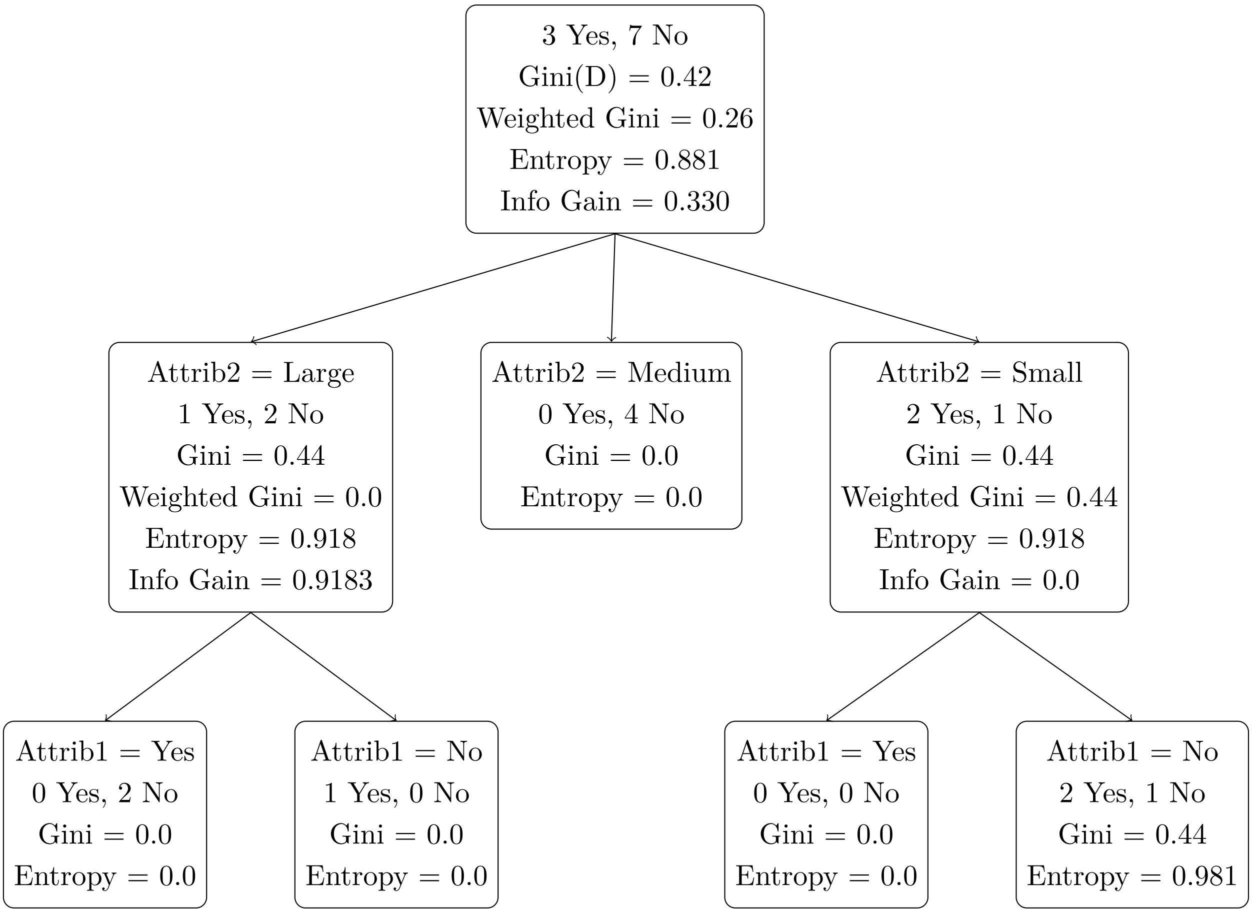
Train Set :

TID	Attrib1	Attrib2	Class
1	Yes	Large	No
2	No	Medium	No
3	No	Small	No
4	Yes	Medium	No
5	No	Large	Yes
6	No	Medium	No
7	Yes	Large	No
8	No	Small	Yes
9	No	Medium	No
10	No	Small	Yes

If Attrib1 is the splitting node:



If Attrib2 is the splitting node:



Test Set

TID	Attrib1	Attrib2	Class
11	No	Small	Yes
12	Yes	Medium	No
13	Yes	Large	No
14	No	Small	Yes
15	No	Large	Yes

Question 2. Handling Numerical Attributes

Original Dataset:

Outlook	Temperature	Humidity	Wind	Play Tennis
Rainy	Cool	59	Strong	No
Sunny	Cool	68	Weak	Yes
Sunny	Mild	72	Strong	Yes
Overcast	Hot	74	Weak	Yes
Overcast	Cool	77	Strong	Yes
Rainy	Cool	79	Weak	Yes
Rainy	Mild	80	Weak	Yes
Sunny	Hot	87	Strong	No
Rainy	Mild	89	Weak	Yes
Sunny	Hot	90	Weak	No
Sunny	Mild	91	Weak	No
Overcast	Hot	93	Weak	Yes
Overcast	Mild	96	Strong	Yes
Rainy	Mild	97	Strong	No

Sorting Data by Numerical Feature:

Outlook	Temperature	Humidity	Wind	Play Tennis
Rainy	Cool	59	Strong	No
Sunny	Cool	68	Weak	Yes
Sunny	Mild	72	Strong	Yes
Overcast	Hot	74	Weak	Yes
Overcast	Cool	77	Strong	Yes
Rainy	Cool	79	Weak	Yes
Rainy	Mild	80	Weak	Yes
Sunny	Hot	87	Strong	No
Rainy	Mild	89	Weak	Yes
Sunny	Hot	90	Weak	No
Sunny	Mild	91	Weak	No
Overcast	Hot	93	Weak	Yes
Overcast	Mild	96	Strong	Yes
Rainy	Mild	97	Strong	No

Calculating Midpoints for Splitting:

Outlook	Temperature	Humidity	Wind	Play Tennis	Mean of consecutive pairs
Rainy	Cool	59	Strong	No	
Sunny	Cool	68	Weak	Yes	63.5
Sunny	Mild	72	Strong	Yes	70
Overcast	Hot	74	Weak	Yes	73
Overcast	Cool	77	Strong	Yes	75.5
Rainy	Cool	79	Weak	Yes	78
Rainy	Mild	80	Weak	Yes	79.5
Sunny	Hot	87	Strong	No	83.5
Rainy	Mild	89	Weak	Yes	88
Sunny	Hot	90	Weak	No	89.5
Sunny	Mild	91	Weak	No	90.5
Overcast	Hot	93	Weak	Yes	92
Overcast	Mild	96	Strong	Yes	94.5
Rainy	Mild	97	Strong	No	96.5

Calculating Information Gain for Numerical Attribute:

$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot Entropy(D_v)$$

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Outlook	Temperature	Humidity	Wind	Play Tennis	Mean	Information Gain
Rainy	Cool	59	Strong	No		
Sunny	Cool	68	Weak	Yes	63.5	0.113
Sunny	Mild	72	Strong	Yes	70	0.01
Overcast	Hot	74	Weak	Yes	73	0.0004
Overcast	Cool	77	Strong	Yes	75.5	0.015
Rainy	Cool	79	Weak	Yes	78	0.0045
Rainy	Mild	80	Weak	Yes	79.5	0.09
Sunny	Hot	87	Strong	No	83.5	**0.152**
Rainy	Mild	89	Weak	Yes	88	0.048
Sunny	Hot	90	Weak	No	89.5	0.102
Sunny	Mild	91	Weak	No	90.5	0.025
Overcast	Hot	94	Weak	Yes	92.5	0.0004
Overcast	Mild	96	Strong	Yes	95	0.01
Rainy	Mild	97	Strong	No	96.5	0.113

Choosing Splitting Point: As seen, **83.5** has the highest information gain, making it the chosen splitting point for the attribute.

Treating "Humidity" as a Categorical Column:

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	> 83.5	Weak	No
Sunny	Hot	> 83.5	Strong	No
Overcast	Hot	> 83.5	Weak	Yes
Rainy	Mild	> 83.5	Weak	Yes
Rainy	Cool	≤ 83.5	Weak	Yes
Rainy	Cool	≤ 83.5	Strong	No
Overcast	Cool	≤ 83.5	Strong	Yes
Sunny	Mild	> 83.5	Weak	No
Sunny	Cool	≤ 83.5	Weak	Yes
Rainy	Mild	≤ 83.5	Weak	Yes
Sunny	Mild	≤ 83.5	Strong	Yes
Overcast	Mild	> 83.5	Strong	Yes
Overcast	Hot	≤ 83.5	Weak	Yes
Rainy	Mild	> 83.5	Strong	No

Question 3. Building Decision Tree cont. Q2

Gini Impurity of the Entire Dataset Given the dataset, we calculate the Gini impurity as follows:

$$\begin{aligned} Gini(D) &= 1 - (p_{\text{Yes}})^2 - (p_{\text{No}})^2 \\ &= 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 \\ &= 0.459 \end{aligned}$$

Calculating Gini Impurity for Each Attribute The Gini impurity for each category of each attribute is calculated, and then the weighted Gini impurity for each attribute is computed.

Outlook Gini impurities for each category:

$$Gini(\text{Sunny}) = 0.48$$
$$Gini(\text{Overcast}) = 0.0$$
$$Gini(\text{Rainy}) = 0.48$$

Weighted Gini impurity:

$$\begin{aligned} Gini(\text{Outlook}) &= \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 \\ &= 0.343 \end{aligned}$$

Temperature Gini impurities for each category:

$$Gini(\text{Hot}) = 0.5$$
$$Gini(\text{Mild}) = 0.444$$
$$Gini(\text{Cool}) = 0.375$$

Weighted Gini impurity:

$$\begin{aligned} Gini(\text{Temperature}) &= \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.444 + \frac{4}{14} \times 0.375 \\ &= 0.440 \end{aligned}$$

Humidity Gini impurities for each category:

$$\begin{aligned} Gini(> 83.5) &= 0.490 \\ Gini(83.5) &= 0.245 \end{aligned}$$

Weighted Gini impurity:

$$\begin{aligned} Gini(\text{Humidity}) &= \frac{7}{14} \times 0.490 + \frac{7}{14} \times 0.245 \\ &= 0.367 \end{aligned}$$

Wind Gini impurities for each category:

$$\begin{aligned} Gini(\text{Weak}) &= 0.375 \\ Gini(\text{Strong}) &= 0.5 \end{aligned}$$

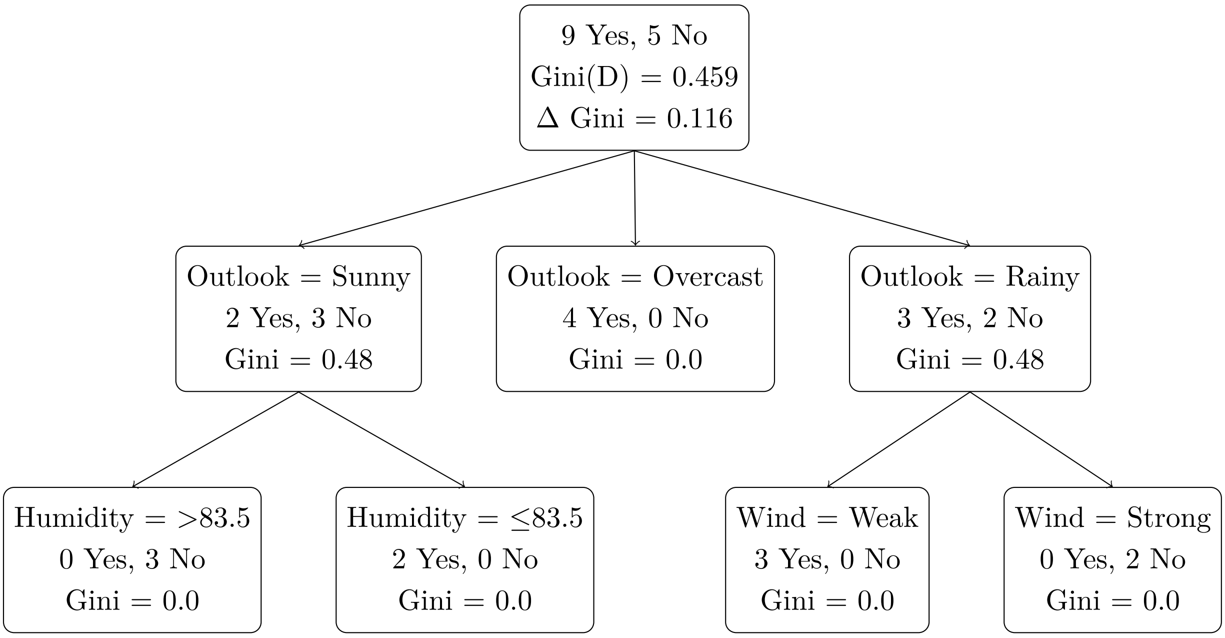
Weighted Gini impurity:

$$\begin{aligned} Gini(\text{Wind}) &= \frac{8}{14} \times 0.375 + \frac{6}{14} \times 0.5 \\ &= 0.429 \end{aligned}$$

Gini Decrease for Each Attribute The Gini decrease for each attribute is calculated to decide the splitting node.

$$\begin{aligned} \Delta Gini(\text{Outlook}) &= 0.459 - 0.343 = 0.116 \\ \Delta Gini(\text{Temperature}) &= 0.459 - 0.440 = 0.019 \\ \Delta Gini(\text{Humidity}) &= 0.459 - 0.367 = 0.092 \\ \Delta Gini(\text{Wind}) &= 0.459 - 0.429 = 0.031 \end{aligned}$$

Decision for Splitting Node Based on the calculated Gini decreases, 'Outlook' is chosen as the splitting node for its highest decrease in Gini impurity.



The prediction for the sample (Sunny, Mild, 85, Weak) is: No