Machine Learning 1

Homework 10: Decision Tree

Student: Nguyễn Thị Minh Ngọc - ID: 11219280

1 Problem 1

Given the training dataset.

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

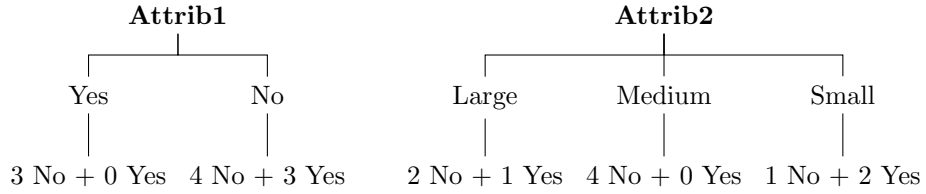
$\longrightarrow$ Test set

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

Requirements

1. Using Gini Impurity to construct the decision tree, then using this decision tree to make predictions for the test set.
2. Do the same but using Information Gain (Entropy)

Solution.



1. Gini Impurity

$$gini(X) = 1 - \left(\frac{3}{10}\right)^2 - \left(\frac{7}{10}\right)^2 = 0.42$$

$$gini(X_{Attrib1=Yes}) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$gini(X_{Attrib1=No}) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = \frac{24}{49} \approx 0.49$$

$$\rightarrow \Delta gini(X, Attrib1) = 0.42 - \frac{3}{10} \cdot 0 - \frac{7}{10} \cdot 0.49 = 0.077$$

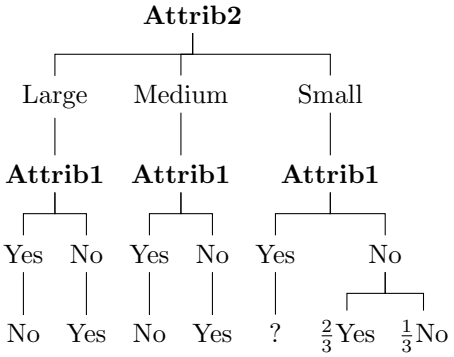
$$gini(X_{Attrib2=Large}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} \approx 0.44$$

$$gini(X_{Attrib2=Medium}) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$gini(X_{Attrib2=Small}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9} \approx 0.44$$

$$\rightarrow \Delta gini(X, Attrib2) = 0.42 - \frac{3}{10} \cdot 0.44 - \frac{4}{10} \cdot 0 - \frac{3}{10} \cdot 0.44 = 0.156$$

Because $\Delta gini(X, Attrib1) < \Delta gini(X, Attrib2)$, we will choose Attrib2 to be the first attribute for splitting the training set. Constructing the decision tree:



$\Longrightarrow$ Making predictions:

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

2. Information Gain (Entropy)

$$entropy(X) = -p_{yes} log_2 p_{yes} - p_{no} log_2 p_{no} = -\frac{3}{10} log_2 \frac{3}{10} - \frac{7}{10} log_2 \frac{7}{10} \approx 0.88$$

$$entropy(X_{Attrib1=Yes}) = -\frac{0}{3} log_2 \frac{0}{3} - \frac{3}{3} log_2 \frac{3}{3} = 0$$

$$entropy(X_{Attrib1=No}) = -\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7} \approx 0.985$$

$$\rightarrow gain(X, Attrib1) = 0.88 - \frac{3}{10} \cdot 0 - \frac{7}{10} \cdot 0.985 = 0.1905$$

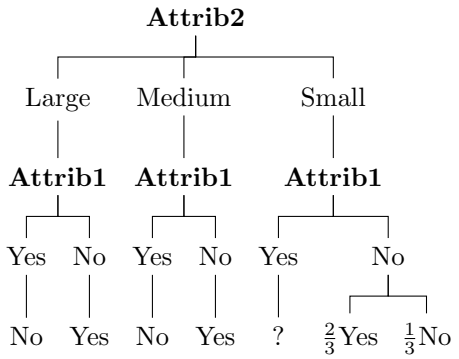
$$entropy(X_{Attrib2=Large}) = -\frac{1}{3} log_2 \frac{1}{3} - \frac{2}{3} log_2 \frac{2}{3} \approx 0.918$$

$$entropy(X_{Attrib2=Medium}) = -\frac{0}{4} log_2 \frac{0}{4} - \frac{4}{4} log_2 \frac{4}{4} = 0$$

$$entropy(X_{Attrib2=Medium}) = -\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3} \approx 0.918$$

$$\rightarrow gain(X, Attrib2) = 0.88 - \frac{3}{10} \cdot 0.918 - \frac{4}{10} \cdot 0 - \frac{3}{10} \cdot 0.918 = 0.3292$$

Because $gain(X, Attrib1) < gain(X, Attrib2)$, we will choose Attrib2 to be the first attribute for splitting the training set. Constructing the decision tree:



$\Longrightarrow$ Making predictions:

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

2 Problem 2

Handling numerical attributes.

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

Solution.

We need to convert the numerical attribute into categorical attribute by finding the best splitting value. To do this, we need to sort the value of the Humidity column first, then calculate the mean of each consecutive pair.

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

$\xrightarrow{\text{sort}}$

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

$\xrightarrow{\text{mean}}$

Candidate
$(59+68)/2$
$(68+72)/2$
$(72+74)/2$
$(74+77)/2$
$(77+79)/2$
$(79+80)/2$
$(80+87)/2$
$(87+89)/2$
$(89+90)/2$
$(90+91)/2$
$(91+93)/2$
$(93+96)/2$
$(96+97)/2$

$\longrightarrow$

Candidate
63.5
70
73
75.5
78
79.5
83.5
88
89.5
90.5
92
94.5
97.5

Then we will calculate the information gain of each splitting value to find the best one based on the following formula:

$$gain(X, a, t) = entropy(X) - \frac{|X_{a \leqslant t}|}{|X|} entropy(X_{a \leqslant t}) - \frac{|X_{a > t}|}{|X|} entropy(X_{a > t})$$

where t is the splitting value and a is the value of Humidity attribute. Taking $gain(X, humidity, 75.5)$ as an example:

$$\begin{aligned}
&entropy(X) = -\frac{5}{14} log_2 \frac{5}{14} - \frac{9}{14} log_2 \frac{9}{14} \approx 0.94 \\
&entropy(X_{a \leqslant 75.5}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} \approx 0.811 \\
&entropy(X_{a \geqslant 75.5}) = -\frac{6}{10} log_2 \frac{6}{10} - \frac{4}{10} log_2 \frac{4}{10} \approx 0.97 \\
&\implies gain(X, humidity, 75.5) = 0.94 - \frac{4}{14} \cdot 0.811 - \frac{10}{14} \cdot 0.97 = 0.015
\end{aligned}$$

Do the same with other values, we get the information gain of each splitting value:

Candidate	Information Gain
63.5	0.113
70	0.01
73	0.0002
75.5	0.015
78	0.045
79.5	0.09
83.5	0.152
88	0.048
89.5	0.101
90.5	0.025
92	0.0002
94.5	0.01
97.5	0.113

Because $gain(X, humidity, 83.5)$ is the highest, 83.5 is the best splitting value. Now we can treat Humidity attribute as a categorical attribute with two possible value: $\leqslant 83.5$ and > 83.5.

3 Problem 3

Using Gini to build the decision tree for data in Problem 2.

Solution

After converting Humidity into categorical attribute, the data become:

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	$\leqslant 83.5$	Weak	Yes
Rainy	Cool	$\leqslant 83.5$	Strong	No
Overcast	Cool	$\leqslant 83.5$	Strong	Yes
Sunny	Mild	> 83.5	Weak	No
Sunny	Cool	$\leqslant 83.5$	Weak	Yes
Rainy	Mild	$\leqslant 83.5$	Weak	Yes
Sunny	Mild	$\leqslant 83.5$	Strong	Yes
Overcast	Mild	> 83.5	Strong	Yes
Overcast	Hot	$\leqslant 83.5$	Weak	Yes
Rainy	Mild	> 83.5	Strong	No

$$gini(X) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = \frac{45}{98} \approx 0.459$$

$$gini(X_{\text{Outlook=Sunny}}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25} = 0.48$$

$$gini(X_{Outlook=Overcast}) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$gini(X_{Outlook=Rainy}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25} = 0.48$$

$$\rightarrow \Delta gini(X, Outlook) = 0.459 - \frac{5}{14} \cdot 0.48 - \frac{4}{14} \cdot 0 - \frac{5}{14} \cdot 0.48 = 0.116$$

Do the same thing with other attributes, we could obtain the following result:

Attributes	$\Delta gini(X, \text{attribute})$
Outlook	0.116
Temperature	0.0185
Humidity	0.092
Wind	0.03

We will choose Outlook attribute as the first splitting attribute as it could bring highest impurity decrease.

- $gini(Sunny) = 0.48$
    - $\Delta gini(Sunny, Temperature) = 0.28$
    - $\Delta gini(Sunny, Humidity) = 0.48$
    - $\Delta gini(Sunny, Wind) = 0.013$
- $gini(Rainy) = 0.48$
    - $\Delta gini(Rainy, Temperature) = 0.013$
    - $\Delta gini(Rainy, Humidity) = 0.013$
    - $\Delta gini(Rainy, Wind) = 0.48$
- $gini(Overcast) = 0$

Hence:

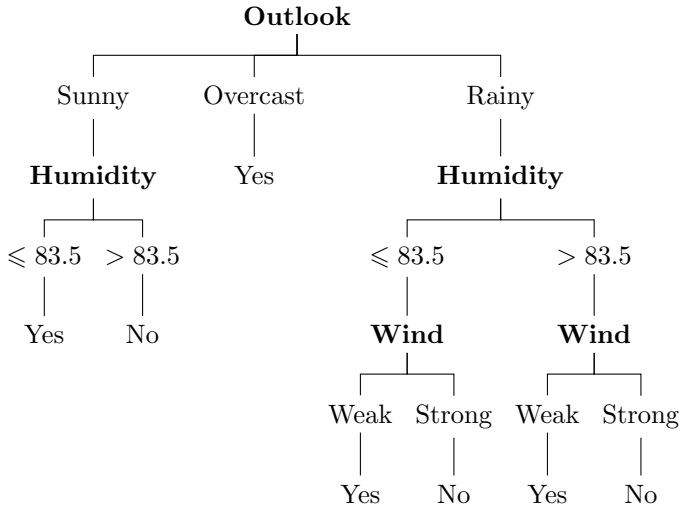
- $\Sigma\Delta gini(Outlook, Temperature) = 0.293$
- $\Sigma\Delta gini(Outlook, Humidity) = 0.493$
- $\Sigma\Delta gini(Outlook, Wind) = 0.493$

Option 1: Choosing Humidity:

- $gini(Sunny, \leqslant 83.5) = 0$
- $gini(Sunny, > 83.5) = 0$
- $gini(Rainy, \leqslant 83.5) = 0.444$
    - $\Delta gini(Rainy, \leqslant 83.5, Temperature) = 0.11$
    - $\Delta gini(Rainy, \leqslant 83.5, Wind) = 0.444$
- $gini(Rainy, > 83.5) = 0.5$
    - $\Delta gini(Rainy, > 83.5, Temperature) = 0$
    - $\Delta gini(Rainy, > 83.5, Wind) = 0.5$

Then:

- $\Sigma\Delta gini(Outlook, Humidity, Temperature) = 0.11$
- $\Sigma\Delta gini(Outlook, Humidity, Wind) = 0.944$



Sample: (Sunny, mild, 85, weak) $\longrightarrow$ No

Option 2: Choosing Wind:

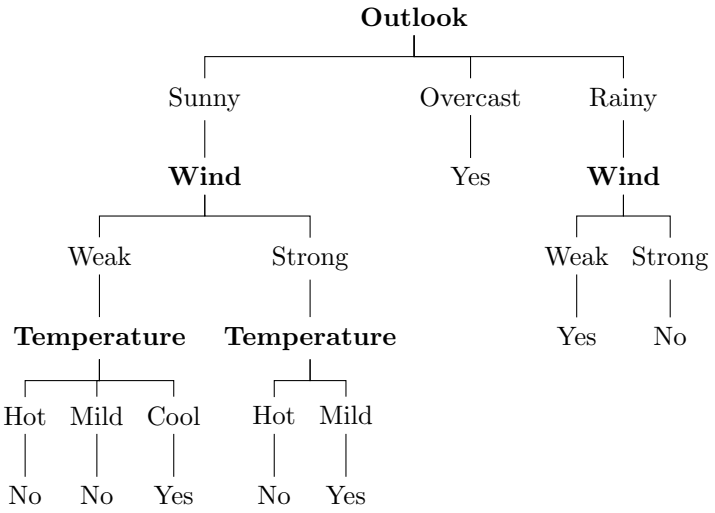
- $gini(Rainy, Weak) = 0$
- $gini(Rainy, Strong) = 0$
- $gini(Sunny, Weak) = 0.444$
    * $\Delta gini(Sunny, Weak, Temperature) = 0.444$
    * $\Delta gini(Sunny, Weak, Humidity) = 0.444$
- $gini(Sunny, > Strong) = 0.5$
    * $\Delta gini(Sunny, Strong, Temperature) = 0.5$
    * $\Delta gini(Sunny, Strong, Wind) = 0.5$

Then:

- $\Sigma\Delta gini(Outlook, Humidity, Temperature) = 0.944$
- $\Sigma\Delta gini(Outlook, Humidity, Wind) = 0.944$

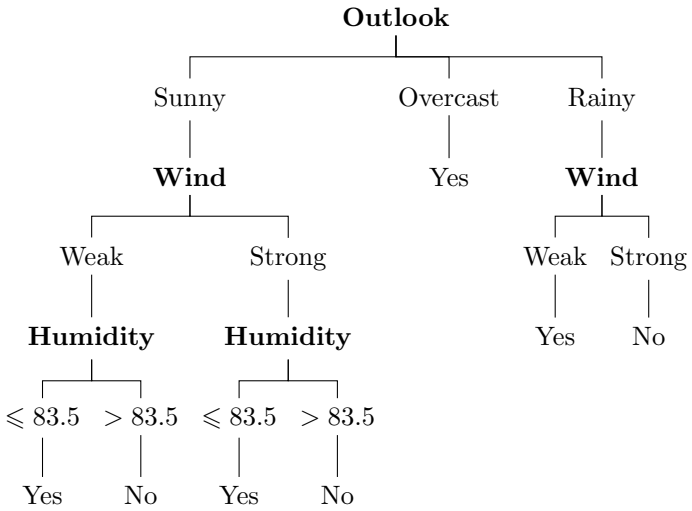
We can choose either Temperature or Humidity.

* Temperature



Sample: (Sunny, mild, 85, weak) $\longrightarrow$ No

* Humidity



Sample: (Sunny, mild, 85, weak) $\longrightarrow$ No