Machine Learning 1

Homework Week 10 - Decision Tree

Tran Hai Nam - 11219279 - DSEB 63

1 Problem 1:

Training dataset

Tid	Attribl	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

Approach 1: Gini Impurty

$$N_{(Yes)} = 3, N_{(No)} = 7$$
$$p_{(Yes)} = \frac{3}{10}, p_{(No)} = \frac{7}{10}$$

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

$\textbf{Gini}(\mathrm{X}_{att1} = yes) = 1 - \left(\frac{3}{3}\right)^2 = 0$

(3 yes - yes)

$\textbf{Gini}(\mathrm{X}_{att1} = no) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.489$

(4 no - no, 3 no - yes)

$$\begin{aligned}\Delta\textbf{Gini}(X,\ \text{atrrib1}) &= \textbf{Gini}(\mathbf{X}) - \frac{7}{10}\cdot\textbf{Gini}(\mathbf{X}_{att1}=yes) - \frac{3}{10}\cdot\textbf{Gini}(\mathbf{X}_{att1}=no)\\ &= 0.42 - \frac{7}{10}\cdot 0.489 - \frac{3}{10}\cdot 0\\ &= 0.0777\end{aligned}$$

And :

$$\textbf{Gini}\left(X_{\text{att2}} =\ \text{large }\right) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.444$$

(1 large - yes , 2 large - no)

$$\textbf{Gini}(X_{\text{att2}} =\ \text{medium }) = 1 - \left(\tfrac{4}{4}\right)^2 - \left(\tfrac{0}{4}\right)^2 = 0$$

(4 medium - no)

$$\textbf{Gini}(X_{\text{att2}} =\ \text{small }) = 1 - \left(\tfrac{2}{3}\right)^2 - \left(\tfrac{1}{3}\right)^2 = 0.444$$

(2 small - yes , 1 small - no)

$$\begin{aligned}\Delta\textbf{Gini}(X,\ \text{atrrib2}) &= \textbf{Gini}(\mathbf{X}) - \frac{3}{10}\cdot\textbf{Gini}(\mathbf{X}_{att2}=large) - \frac{4}{10}\cdot\textbf{Gini}(\mathbf{X}_{att2}=medium)\\ &\quad - \frac{3}{10}\cdot\textbf{Gini}(\mathbf{X}_{att2}=small)\\ &= 0.42 - \frac{3}{10}\cdot 0.444 - \frac{4}{10}\cdot 0 - \frac{3}{10}\cdot 0.444\\ &= 0.402\end{aligned}$$

Since $\Delta\textbf{Gini}(X,\ \text{atrrib2}) > \Delta\textbf{Gini}(X,\ \text{atrrib1})$

$\longrightarrow$ **We choose Attrb2 to be the root.**

Approach 2: Information gain
(best attribute = highest information gain)

$$\text{Entropy }(X) = -\left(\frac{7}{10}\right)\log_2\left(\frac{7}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{40}\right) = 0.88129$$

$$\text{Entropy }(X,\ \text{att1} = \text{yes }) = -\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) = 0$$

(3 yes - no)

$$\text{Entropy}\,(X,\ \text{att1} = \text{no }) = -\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) = 0.985$$

(4 no - no , 3 no - yes)

$$\begin{aligned}\textbf{Gain}(X, attrb1) &= 0.88129 - \frac{3}{10}\cdot 0 - \frac{7}{10}\cdot 0.985 \\ &= 0.19179\end{aligned}$$

$$\text{Entropy }(X,\ \text{att2} = \text{large }) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9182$$

(2 large - no , 1 large - yes)

$$\text{Entropy }(X, \text{att2} = \text{small }) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.9182$$

(2 small - yes, 1 small - no)

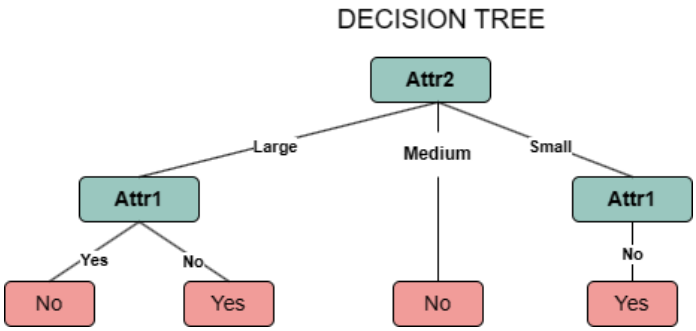
$$\text{Entropy }(X, \text{att2} = \text{medium }) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) = 0$$

(4 medium - no)

$$\textbf{Gain}(X, att2) = 0.88129 - \frac{3}{10}\cdot 0.9182 - \frac{4}{10}\cdot 0 - \frac{3}{10}\cdot 0.9182 = 0.805$$

Since $\textbf{Gain}(X,\ \text{atrrib2}) > \textbf{Gain}(X,\ \text{atrrib1})$

$\longrightarrow$ **We choose Attrb2 to be the root.**



2 Problem 2:

Outlook	Temperature	Humidity	Wind	Play Tennis?
Sunny	Hot	High	Weak	**No**
Sunny	Hot	High	Strong	**No**
Overcast	Hot	High	Weak	**Yes**
Rainy	Mild	High	Weak	**Yes**
Rainy	Cool	Normal	Weak	**Yes**
Rainy	Cool	Normal	Strong	**No**
Overcast	Cool	Normal	Strong	**Yes**
Sunny	Mild	High	Weak	**No**
Sunny	Cool	Normal	Weak	**Yes**
Rainy	Mild	Normal	Weak	**Yes**
Sunny	Mild	Normal	Strong	**Yes**
Overcast	Mild	High	Strong	**Yes**
Overcast	Hot	Normal	Weak	**Yes**
Rainy	Mild	High	Strong	**No**

Gain of numerical attribute a if we split at value t

$$\textbf{gain}(X, a, t) = \textbf{entropy}(X) - \frac{|X_{a\leq t}|}{|X|}\textbf{entropy}\left(X_{a\leq t}\right) - \frac{|X_a > t|}{|X|}\textbf{entropy}\left(X_{a>t}\right)$$

$\rightarrow$ **Sort**

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**

Sort

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

Mean of each consecutive pair

Candidate split values
63
70
73
75.5
78
79.5
83.5
88
89.5
90.5
92
94.5
96.5

$$\textbf{entropy}\ (X) = -\tfrac{5}{14}\log_2\left(\tfrac{5}{14}\right) - \tfrac{9}{14}\log_2\left(\tfrac{9}{14}\right) = 0.94$$

$$\textbf{entropy}(X_a \leq 83.5) = -\tfrac{6}{7}\log_2\left(\tfrac{6}{7}\right) - \tfrac{1}{7}\log_2\left(\tfrac{6}{7}\right) = 0.59$$

(6 yes - 1 no)

$$\textbf{entropy}\ (X_a > 83.5) = -\tfrac{4}{7}\log_2\left(\tfrac{4}{7}\right) - \tfrac{3}{7}\log_2\left(\tfrac{3}{7}\right) = 0.985$$

$$\textbf{gain}(x,\ \text{humidity}, 83.5) = 0.94 - \tfrac{7}{14}\cdot 0.59 - \tfrac{7}{14}\cdot 0.985 = 0.1525$$

3 Problem 3:

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**

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

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

$$\textbf{gini}(\text{ Outlook } = \text{ Rainy }) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48.$$

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

$$\Delta \textbf{ gini}\,(X, \text{ outlook }) = 0.459 - \frac{5}{14} \cdot 0.48 - \frac{5}{14} \cdot 0.48 = 0.1161$$

$\rightarrow$ **best attribute**