Decision Trees

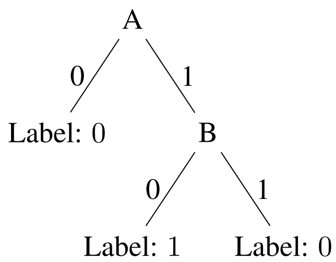
Exercise 1 : Decision Trees

Construct by hand decision trees corresponding to each of the following Boolean formulas. The examples $(\mathbf{x}, c) \in D$ consist of a feature vector $\mathbf{x}$ where each component corresponds to one of the Boolean variables $(A, B, \dots)$ used in the formula, and each example corresponds to one interpretation (i.e. assignment of 0/1 to the Boolean variables). The target concept c is the truth value of the formula given that interpretation. Assume the set D contains examples with all possible combinations of attribute values.

Hint: It may be helpful to write out the set D for each formula as a truth table.

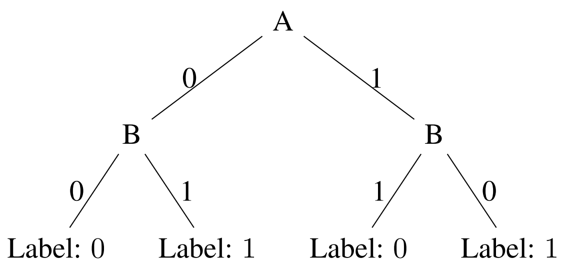
(a) $A \wedge \neg B$

Answer



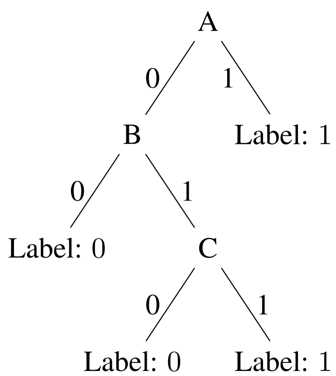
(b) A *XOR* B

Answer



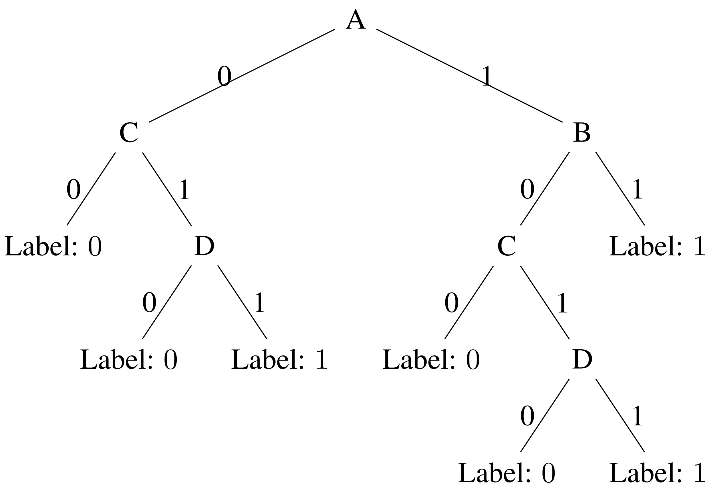
(c) $A \vee (B \wedge C)$

Answer



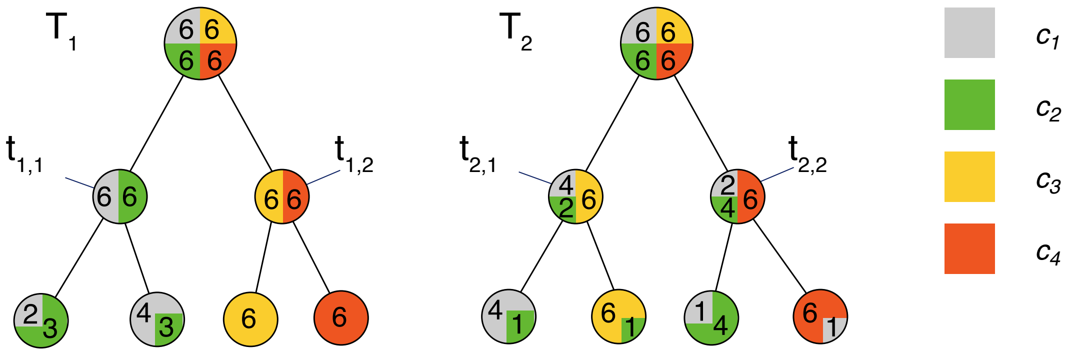
(d) $(A \wedge B) \vee (C \wedge D)$

Answer



Exercise 2 : Impurity Functions

Let D be a set of examples over a feature space $\mathbf{X}$ and a set of classes $C = \{c_1, c_2, c_3, c_4\}$, with $|D| = 24$. Consider the following illustration of two possible decision trees, T_1 and T_2 – the colors represent the classes present in each subset $D(t_i)$ represented by node $t_{i,j}$ of T_i; the numbers denote how many examples of each class are present.



(a) First, consider only the first split that each of the two trees makes: compute $\Delta\iota(D, \{D(t_{1,1}), D(t_{1,2})\})$ and $\Delta\iota(D, \{D(t_{2,1}), D(t_{2,2})\})$ with (1) the misclassification rate $\iota_{misclass}$ and (2) the entropy criterion $\iota_{entropy}$ as splitting criterion.

Interpret the results: which of $\{D(t_{1,1}), D(t_{1,2})\}$ or $\{D(t_{2,1}), D(t_{2,2})\}$ is the better first split?

Answer

$$\begin{aligned}
&\Delta\iota_{\textit{misclass}}(D, \{D(t_{1,1}), D(t_{1,2})\}) \\
&= \iota_{\textit{misclass}}(D) - \textstyle\sum_{l=1}^{2} \frac{|D(t_{1,l})|}{|D|} \cdot \iota_{\textit{misclass}}(D(t_{1,l})) \\
&= (1 - \max\{\tfrac{6}{24}, \tfrac{6}{24}, \tfrac{6}{24}, \tfrac{6}{24}\}) - 2 \cdot \tfrac{12}{24} \cdot (1 - \max\{\tfrac{6}{12}, \tfrac{6}{12}\}) \\
&= (1 - 0.25) - 2 \cdot 0.5 \cdot (1 - 0.5) \\
&= 0.75 - 2 \cdot 0.5 \cdot 0.5 \\
&= 0.25 \\
&\Delta\iota_{\textit{misclass}}(D, \{D(t_{2,1}), D(t_{2,2})\}) \\
&= \iota_{\textit{misclass}}(D) - \textstyle\sum_{l=1}^{2} \frac{|D(t_{2,l})|}{|D|} \cdot \iota_{\textit{misclass}}(D(t_{2,l})) \\
&= (1 - \max\{\tfrac{6}{24}, \tfrac{6}{24}, \tfrac{6}{24}, \tfrac{6}{24}\}) - 2 \cdot \tfrac{12}{24} \cdot (1 - \max\{\tfrac{6}{12}, \tfrac{4}{12}, \tfrac{2}{12}\}) \\
&= (1 - 0.25) - 2 \cdot 0.5 \cdot (1 - 0.5) \\
&= 0.75 - 2 \cdot 0.5 \cdot 0.5 \\
&= 0.25 \\
&\Delta\iota_{\textit{entropy}}(D, \{D(t_{1,1}), D(t_{1,2})\}) \\
&= \iota_{\textit{entropy}}(D) - \textstyle\sum_{l=1}^{2} \frac{|D(t_{1,l})|}{|D|} \cdot \iota_{\textit{entropy}}(D(t_{1,l})) \\
&= -4 \cdot \left(\tfrac{6}{24} \log_2 \tfrac{6}{24}\right) - 2 \cdot \tfrac{12}{24} \cdot \left(-\left(\tfrac{6}{12} \cdot \log_2 \tfrac{6}{12}\right) - \left(\tfrac{6}{12} \cdot \log_2 \tfrac{6}{12}\right)\right) \\
&= -4 \cdot (-0.5) - 2 \cdot 0.5 \cdot (0.5 + 0.5) \\
&= 1 \\
&\Delta\iota_{\textit{entropy}}(D, \{D(t_{2,1}), D(t_{2,2})\}) \\
&= \iota_{\textit{entropy}}(D) - \textstyle\sum_{l=1}^{2} \frac{|D(t_{2,l})|}{|D|} \cdot \iota_{\textit{entropy}}(D(t_{2,l})) \\
&= -4 \cdot \left(\tfrac{6}{24} \log_2 \tfrac{6}{24}\right) - 2 \cdot \tfrac{12}{24} \cdot \left(-\left(\tfrac{6}{12} \cdot \log_2 \tfrac{6}{12}\right) - \left(\tfrac{4}{12} \cdot \log_2 \tfrac{4}{12}\right) - \left(\tfrac{2}{12} \cdot \log_2 \tfrac{2}{12}\right)\right) \\
&= -4 \cdot (-0.5) - 2 \cdot 0.5 \cdot (0.5 + 0.528 + 0.431) \\
&= 0.541
\end{aligned}$$

With the misclassification rate both splits are identically evaluated. The entropy criterion prefers pure example sets. The split in T_1 gets rated higher. Intuitively, the entropy criterion is right: after the first split in T_1, there is “less work to do” to purify all example sets.

(b) If we compare T_1 and T_2 in terms of their misclassification rate on D, which one is the better decision tree?

Answer

According to the training set error T_2, i.e., $Err(T_2, D) = \frac{4}{24}$, is better than T_1, i.e. $Err(T_1, D) = \frac{5}{24}$.

(c) Assuming the splits shown are the only possibilities, which of T_1 or T_2 would the ID3 algorithm construct, and why?

Answer

ID3 uses information gain (i.e., entropy impurity reduction) as the split criterion. Hence, as the first split, $\{D(t_{1,1}), D(t_{1,2})\}$ would be chosen, and the “less good” decision tree would result; this is because ID3 searches the hypothesis space by greedy local optimization. There is no guarantee to find a globally optimal hypothesis.

Exercise 3 : Decision Trees

Given is the following dataset to classifiy whether a dog is dangerous or well-behaved in character:

Color	Fur	Size	Character (C)
brown	ragged	small	well-behaved
black	ragged	big	dangerous
black	smooth	big	dangerous
black	curly	small	well-behaved
white	curly	small	well-behaved
white	smooth	small	dangerous
red	ragged	big	well-behaved

(a) Use the ID3 algorithm with $\iota_{entropy}$ as the impurity function to determine the tree T.

Answer

- Determine $\iota_{entropy}(D)$:

$$\begin{aligned}\iota_{entropy}(D) &= -\sum_{i=1}^{k} P(A_i) \cdot \log_2 P(A_i) \\ &= -\left[\frac{4}{7} \cdot \log_2 \frac{4}{7} + \frac{3}{7} \cdot \log_2 \frac{3}{7}\right] \\ &\approx 0.985\end{aligned}$$

- Determine $\Delta\iota_{entropy} = 0.985 - \sum_{l=1}^{m} \frac{|D_l|}{|D|} \cdot \iota_{entropy}(D_l)$ for each attribute and choose the attribute with maximum delta (i.e., information gain) to split:
  - Attribute *Color*:

Color	well-behaved	dangerous	Probability
brown	1	0	$P(\textit{brown}) = 1/7$
black	1	2	$P(\textit{black}) = 3/7$
white	1	1	$P(\textit{white}) = 2/7$
red	1	0	$P(\textit{red}) = 1/7$

$$\begin{aligned}\Delta\iota_{entropy} &= 0.985 - \left[\frac{1}{7}\left(-\left(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right)\right) + \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)\right)\right. \\ &\quad \left. + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) + \frac{1}{7}\left(-\left(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right)\right)\right] \\ &= 0.985 - \left[0 + \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)\right) + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) + 0\right] \\ &\approx 0.306\end{aligned}$$

  - Attribute *Fur*:

Fur	well-behaved	dangerous	Probability
ragged	2	1	$P(\textit{ragged}) = 3/7$
smooth	0	2	$P(\textit{smooth}) = 2/7$
curly	2	0	$P(\textit{curly}) = 2/7$

$$\begin{aligned}\Delta\iota_{entropy} &= 0.985 - \left[\frac{3}{7}\left(-\left(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}\right)\right)+\frac{2}{7}\left(-\left(\frac{0}{2}\log_2\frac{0}{2}+\frac{2}{2}\log_2\frac{2}{2}\right)\right)\right. \\ &\quad \left.+\frac{2}{7}\left(-\left(\frac{2}{2}\log_2\frac{2}{2}+\frac{0}{2}\log_2\frac{0}{2}\right)\right)\right] \\ &= 0.985 - \left[\frac{3}{7}\left(-\left(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}\right)\right)+0+0\right] \\ &\approx 0.591\end{aligned}$$

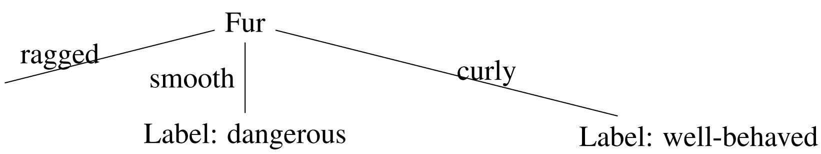
- Attribute *Size*:

Size	well-behaved	dangerous	Probability
small	3	1	$P(\mathit{small}) = 4/7$
big	1	2	$P(\mathit{big}) = 3/7$

$$\begin{aligned}\Delta\iota_{entropy} &= 0.985 - \left[\frac{4}{7}\left(-\left(\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)\right)+\frac{3}{7}\left(-\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)\right)\right] \\ &\approx 0.128\end{aligned}$$

$\Delta\iota_{entropy}$ is maximal for attribute *Fur*. Generated tree with reduced dataset is pictured below.

Color	Size	Character (C)
brown	small	well-behaved
black	big	dangerous
red	big	well-behaved



- ID3 is applied recursively to remaining non-terminal nodes. Determine $\iota_{entropy}(D)$ for the reduced dataset:

$$\begin{aligned}\iota_{entropy}(D) &= -\sum_{i=1}^{k} P(A_i)\cdot\log_2 P(A_i) \\ &= -\left[\frac{1}{3}\cdot\log_2\frac{1}{3}+\frac{2}{3}\cdot\log_2\frac{2}{3}\right] \\ &\approx 0.918\end{aligned}$$

- Determine $\Delta\iota_{entropy} = 0.918 - \sum_{l=1}^{m}\frac{|D_l|}{|D|}\cdot\iota_{entropy}(D_l)$ for each remaining attribute and choose the attribute with maximum delta (i.e., information gain) to split:
  - Attribute *Color*:

Color	well-behaved	dangerous	Probability
brown	1	0	$P(\mathit{brown}) = 1/3$
black	0	1	$P(\mathit{black}) = 1/3$
red	1	0	$P(\mathit{red}) = 1/3$

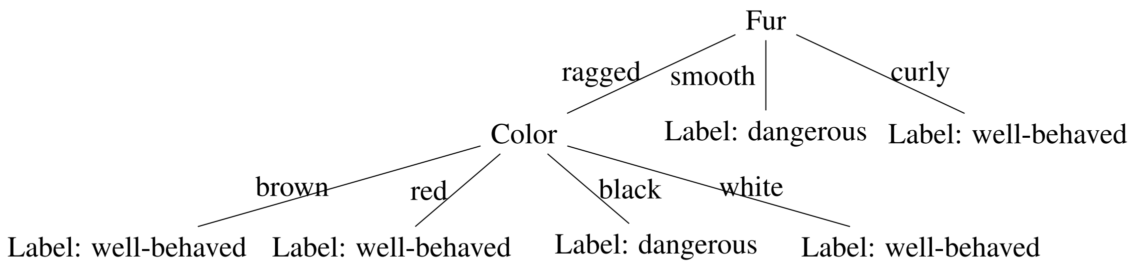
$$\begin{aligned}\Delta\iota_{entropy}(D) &= 0.918 - \left[\frac{1}{3}\left(-\left(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\right)\right)+\frac{1}{3}\left(-\left(\frac{0}{1}\log_2\frac{0}{1}+\frac{1}{1}\log_2\frac{1}{1}\right)\right)\right. \\ &\quad \left.+\frac{1}{3}\left(-\left(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\right)\right)\right] \\ &= 0.918\end{aligned}$$

– Attribute *Size*:

Size	well-behaved	dangerous	Probability
small	1	0	$P(small) = 1/3$
big	1	1	$P(big) = 2/3$

$$\begin{aligned}\Delta\iota_{entropy}(D) &= 0.918 - \left[\frac{1}{3}\left(-\left(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right)\right) + \frac{2}{3}\left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right)\right] \\ &= 0.918 - \left[0 + \frac{2}{3}\left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right)\right] \\ &\approx 0.252\end{aligned}$$

$\Delta\iota_{entropy}$ is maximal for attribute *Color*. As *white* does not occur in the reduced dataset, the most common class of the reduced dataset is chosen as label. Generated tree is pictured below.



(b) Classify the new example (Color=black, Fur=ragged, Size=small) using T.

Answer

1. Check attribute fur.
2. Fur=ragged $\rightarrow$ Check attribute color.
3. color=black $\rightarrow$ Assign class=dangerous