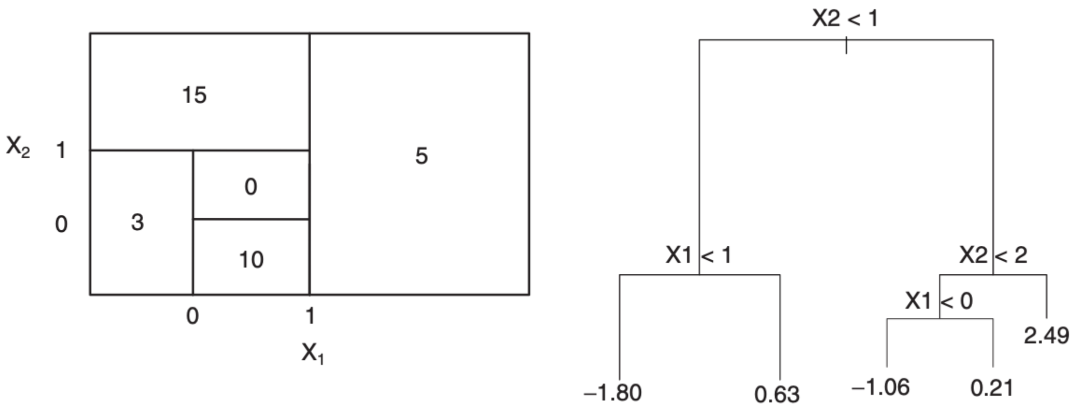
Tree-based methods

Machine Learning 2021-2022 - UMONS
Souhaib Ben Taieb

1

1. Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 1. The numbers inside the boxes indicate the mean of Y within each region.

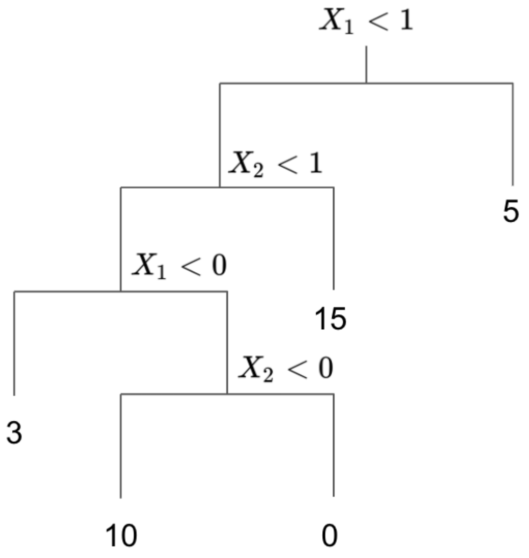
2. Create a diagram similar to the left-hand panel of Figure 1, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.



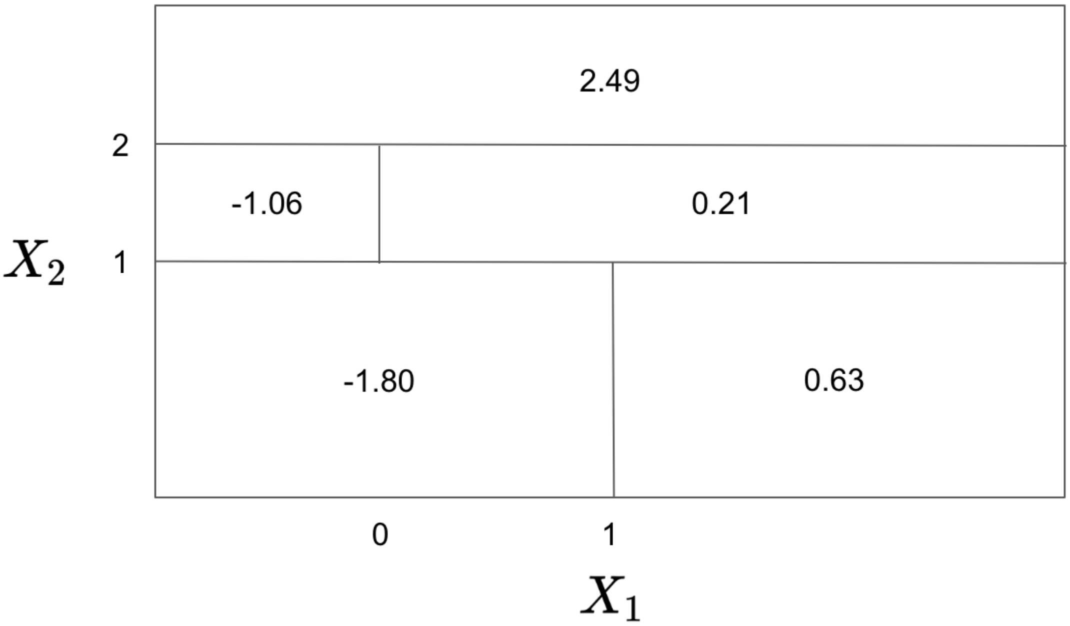
(This question is from ISLR, Section 8.4, exercise 4).

Solution :

1.1.



1.2.



2

Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{Class is Red} \,|X = x)$: 0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

(This question is from ISLR, Section 8.4, exercise 5).

Solution :

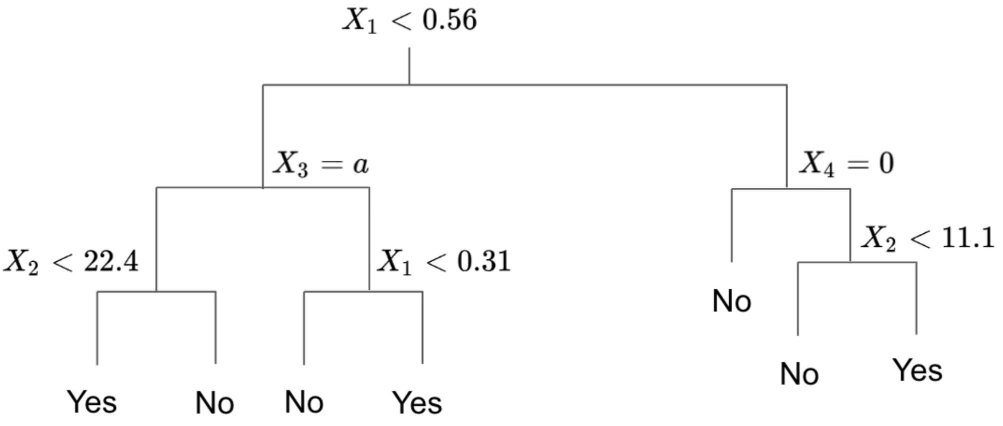
Suppose that the classification threshold is set at 0.5, i.e. the class red is predicted if $P(\text{Class is Red} \,|X = x) \geq 0.5$. Under this threshold, the predicted class for X in each of the bootsrapped samples is : Green, Green, Green, Green, Red, Red, Red, Red, Red, Red. Taking a majority vote, the final predicted class for X is Red.

If we now average the probabilities obtained in each boostrapped samples, we have $\bar{P}(\text{Class is Red} \,|X = x) = 0.45$, and the class predicted for X is now Green.

3

Given the decision tree of Figure [3], how would the following observations be classified ?

X_1	X_2	X_3	X_4	Y
0.48	18.1	a	1	
0.64	32.5	a	0	
0.12	26.5	b	0	
0.69	6.7	c	1	
0.43	18.6	c	0	
0.84	16.5	a	1	
0.33	28.5	a	1	
0.92	6.3	c	1	
0.96	12.1	b	0	
0.16	13.1	b	1	



Solution :

X_1	X_2	X_3	X_4	Y
0.48	18.1	a	1	Yes
0.64	32.5	a	0	No
0.12	26.5	b	0	No
0.69	6.7	c	1	No
0.43	18.6	c	0	Yes
0.84	16.5	a	1	Yes
0.33	28.5	a	1	No
0.92	6.3	c	1	No
0.96	12.1	b	0	No
0.16	13.1	b	1	No

4

Build a classification tree using the information gain for the following dataset:

Size	Orbit	Temperature	Habitable
Big	Far	205	No
Big	Near	205	No
Big	Near	260	Yes
Big	Near	380	Yes
Small	Far	205	No
Small	Far	260	Yes
Small	Near	260	Yes
Small	Near	380	No
Small	Near	380	No

Solution :

We will build our decision tree using the recursive binary splitting algorithm such that, at each split, the information gain is maximized :

$$\mathrm{IG}(Y|X) = H(Y) - H(Y|X)$$

where :

$$H(Y) = -\sum_{y\in\mathcal{Y}} p(y)\log_2 p(y)$$

and :

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) && (1)\\ &= \sum_{x\in\mathcal{X}} p(x)\Big(\sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x)\Big) && (2)\end{aligned}$$

where $Y \in \{Yes, No\}$ indicates whether or not the planet is habitable, and X is a binary attribute constructed from the original attributes (e.g. Size, Orbit, Temperature) which realization corresponds to each sides of the split.

To find out which split leads to the highest information gain, we must account for every variables and for every possible split amongst them. Once the split leading to the highest information gain is found, we must repeat the operation for each of the obtained splits, until a stopping criterion is met. In practice, the stopping criterion is a minimum number of samples per leaf (e.g. 5). However, for this exercise, we will grow the tree to its full depth.

The entropy of Y is equal to:

$$H(Y) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.99$$

1. **First iteration**

1.1 *Variable 'Size'*

Let's begin with the variable 'Size', that we will note $S \in \{Big, Small\}$. We have :

$H(Y|S = Big) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$H(Y|S = Small) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$$\begin{aligned} H(Y|S) &= p(S = Big)H(Y|S = Big) + p(S = Small)H(Y|S = Small)\\ &= \frac{4}{9} * 1 + \frac{5}{9} * 0.97\\ &= 0.98\end{aligned}$$

$$\begin{aligned}\mathrm{IG}(Y|S) &= H(Y) - H(Y|S)\\ &= 0.99 - 0.98 = 0.01\end{aligned}$$

1.2 *Variable 'Orbit'*

Let's note the variable 'Orbit' as $O \in \{Far, Near\}$. We have :
$H(Y|O = Far) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$
$H(Y|O = Near) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$$\begin{aligned}H(Y|O) &= p(O = Far)H(Y|O = Far) + p(O = Near)H(Y|O = Near)\\ &= \frac{3}{9} * 0.92 + \frac{6}{9} * 1\\ &= 0.97\end{aligned}$$

$$\begin{aligned}\mathrm{IG}(Y|O) &= H(Y) - H(Y|0)\\ &= 0.99 - 0.97 = 0.02\end{aligned}$$

1.3 *Variable 'Temperature'*

Let's note the variable 'Temperature' as $T = \{205, 260, 380\}$ from which we can create the binary variables $T_1 \in \{205, \neq 205\}$ and $T_2 \in \{260, \neq 260\}$ and $T_3 \in \{380, \neq 380\}$. For T_1, T_2 and T_3, we have :
$H(Y|T_1 = 205) = 0$
$H(Y|T_1 \neq 205) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.92$

$$\begin{aligned}H(Y|T_1) &= p(T_1 = 205)H(Y|T_1 = 205) + p(T_1 \neq 205)H(Y|T_1 \neq 205)\\ &= \frac{3}{9} * 0 + \frac{6}{9} * 0.92\\ &= 0.61\end{aligned}$$

$$\begin{aligned}\mathrm{IG}(Y|T_1) &= H(Y) - H(Y|T_1)\\ &= 0.99 - 0.61 = 0.38\end{aligned}$$

$H(Y|T_2 = 260) = 0$
$H(Y|T_2 \neq 260) = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} = 0.65$

$$\begin{aligned}H(Y|T_2) &= p(T_2 = 260)H(Y|T_2 = 260) + p(T_2 \neq 260)H(Y|T_1 \neq 260)\\ &= \frac{3}{9} * 0 + \frac{6}{9} * 0.65\\ &= 0.43\end{aligned}$$

$$\begin{aligned}\mathrm{IG}(Y|T_2) &= H(Y) - H(Y|T_2)\\ &= 0.99 - 0.43 = 0.56\end{aligned}$$

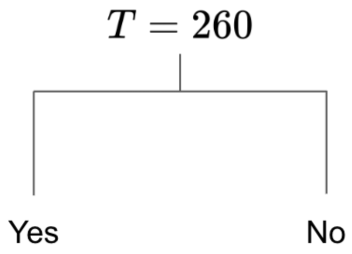
$H(Y|T_3 = 380) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$

$H(Y|T_3 \neq 380) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

$$\begin{aligned}H(Y|T_3) &= p(T_3 = 380)H(Y|T_3 = 380) + p(T_3 \neq 380)H(Y|T_3 \neq 380)\\ &= \frac{3}{9} * 0.92 + \frac{6}{9} * 1\\ &= 0.97\end{aligned}$$

$$\begin{aligned}\mathrm{IG}(Y|T_3) &= H(Y) - H(Y|T_3)\\ &= 0.99 - 0.97 = 0.02\end{aligned}$$

The highest information gain is obtained by splitting the variable 'Temperature' at the value of $T = 260$. This will thus form the first node of our decision tree. By taking a majority vote on the class of the observations that fall in either of the two regions $R_1 = \{T|T = 260\}$ and $R_2 = \{T|T \neq 260\}$, we get the decision tree of Figure [4].



As all observations that fall into the region R_1 all belong to the same class (i.e $Y = Yes$), the entropy is null and there is no point in trying to find another split that would further decreases the entropy. However, we can further split the region R_2, and so begins the second iteration of the algorithm.

2. **Second iteration**

The entropy of the variable Y in R_2 is given by :

$$H(Y|T \neq 260) = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} = 0.65$$

2.1. *Variable 'Size'*

$H(Y|T \neq 260, S = Big) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$

$H(Y|T \neq 260, S = Small) = 0$

$$\begin{aligned}H(Y|T \neq 260, S) &= p(S = Big)H(Y|T \neq 260, S = Big) + p(S = Small)H(Y|T \neq 260, S = Small)\\ &= \frac{3}{6} * 0.92 + \frac{3}{6} * 0\\ &= 0.46\end{aligned}$$

$$\begin{aligned}\text{IG}(Y|T \neq 260, S) &= H(Y|T \neq 260) - H(Y|T \neq 260, S)\\ &= 0.65 - 0.46 = 0.19\end{aligned}$$

2.2. *Variable 'Orbit'*

$H(Y|T \neq 260, O = Far) = 0$

$H(Y|T \neq 260, O = Near) = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4} = 0.81$

$$\begin{aligned}H(Y|T \neq 260, O) &= p(O = Far)H(Y|T \neq 260, O = Far) + p(O = Near)H(Y|T \neq 260, O = Near)\\ &= \frac{2}{6} * 0 + \frac{4}{6} * 0.81\\ &= 0.54\end{aligned}$$

$$\begin{aligned}\text{IG}(Y|T \neq 260, O) &= H(Y|T \neq 260) - H(Y|T \neq 260, O)\\ &= 0.65 - 0.54 = 0.11\end{aligned}$$

2.3 *Variable 'Temperature'*

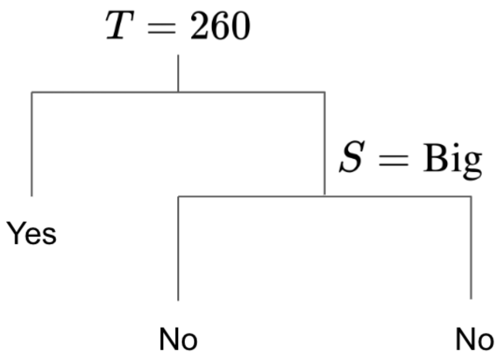
$H(Y|T \neq 260, T = 205) = 0$

$H(Y|T \neq 260, T \neq 205) = -\frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3} = 0.92$

$$\begin{aligned}H(Y|T \neq 260, T) &= p(T = 205)H(Y|T \neq 260, T = 205) + p(T \neq 205)H(Y|T \neq 260, T \neq 205)\\ &= \frac{3}{6} * 0 + \frac{3}{6} * 0.92\\ &= 0.46\end{aligned}$$

$$\begin{aligned}\text{IG}(Y|T \neq 260, T) &= H(Y|T \neq 260) - H(Y|T \neq 260, T)\\ &= 0.65 - 0.46 = 0.19\end{aligned}$$

Here we have a tie between the variables 'Size' and 'Temperature' that both lead to an information gain of 0.19. Let's choose the variable 'Size' to perform our split. We now have three regions, $R_1 = \{T|T = 260\}$, $R_2 = \{T, S|T \neq 260, S = Big\}$ and $R_3 = \{T, S|T \neq 260, S = Small\}$, and by taking a majority vote in each region, we get the decision tree of Figure [4] :



The entropy in R_3 cannot be further minimized, but well in R_2. The third iteration of the algorithm begins.

3. **Third iteration**

The variables that could lead to further splits of the input space are the variables 'Orbit' and 'Temperature'.

The entropy of the variable Y in R_2 is given by :

$$H(Y|T \neq 260, S = Big) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$$

3.1. *Variable 'Orbit'*

$H(Y|T \neq 260, S = Big, O = Far) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$H(Y|T \neq 260, S = Big, O = Near) = 0$

$$\begin{aligned} H(Y|T \neq 260, S = Big, O) &= p(O = Far)H(Y|T \neq 260, S = Big, O = Far) + p(O = Near)H(Y|T \neq 260, S = Big, O = Near) \\ &= \frac{2}{3} * 1 + \frac{1}{3} * 0 \\ &= 0.66 \end{aligned}$$

$$\begin{aligned} \mathrm{IG}(Y|T \neq 260, S = Big, O) &= H(Y|T \neq 260, S = Big) - H(Y|T \neq 260, S = Big, O) \\ &= 0.92 - 0.66 = 0.26 \end{aligned}$$

3.2. *Variable 'Temperature'*

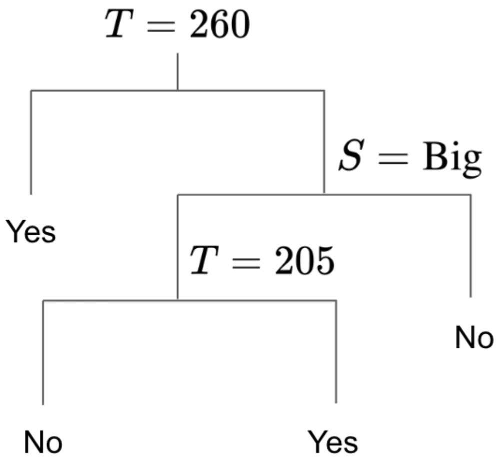
$H(Y|S = Big, T = 205) = 0$

$H(Y|T \neq 260, S = Big, T \neq 205) = 0$

$$\begin{aligned} H(Y|S = Big, T) &= p(T = 205)H(Y|S = Big, T = 205) + p(T \neq 205)H(Y|T \neq 260, S = Big, T \neq 205) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \mathrm{IG}(Y|S = Big, T) &= H(Y|T \neq 260, S = Big) - H(Y|S = Big, T) \\ &= 0.92 - 0 = 0.92 \end{aligned}$$

We split the variable 'Temperature' at the value of $T = 205$, which now leads to four regions : $R_1 = \{T|T = 260\}$, $R_2 = \{T,S|T = 205, S = Big\}$, $R_3 = \{T,S|T = 380, S = Big\}$ and $R_4 = \{T,S|T \neq 260, S = Small\}$. By taking a majority vote on the classes in each region, we get the decision tree of Figure [4] :



We cannot further decrease the entropy in any of the regions R_i, and so the algorithm stops.