Tree-based methods

Machine Learning 2023-2024 - 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.

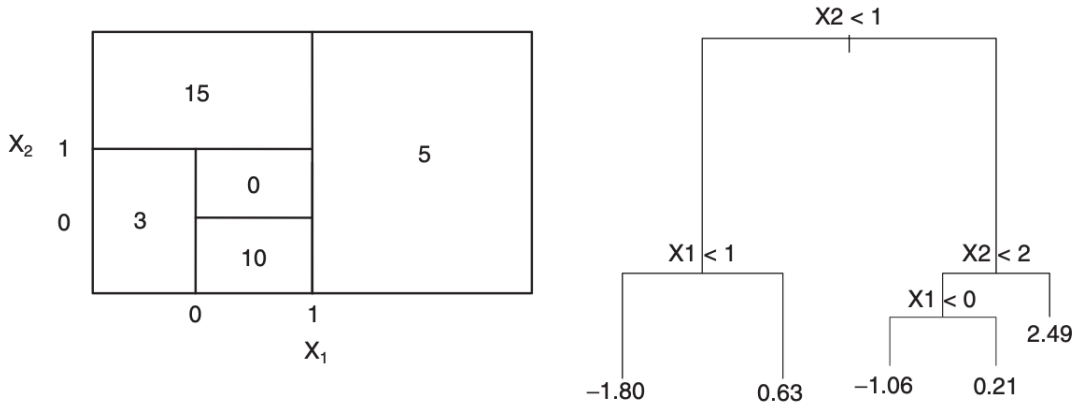
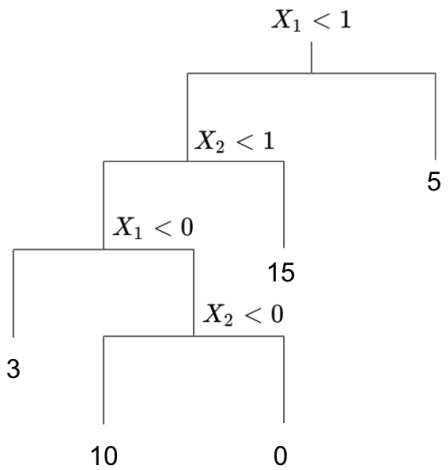


Figure 1:

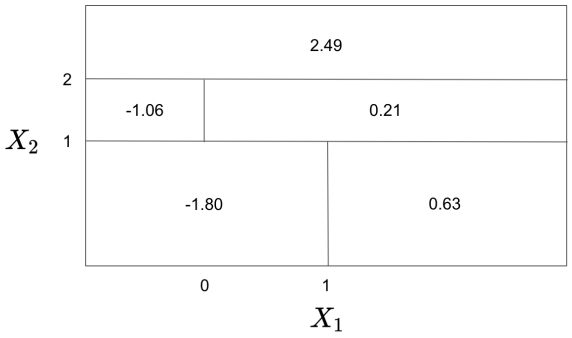
(This question is from ISLR, Section 8.4, exercise 4).

Solution:

1.1.



1.2.



2

Given the decision tree of Figure 2, 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	

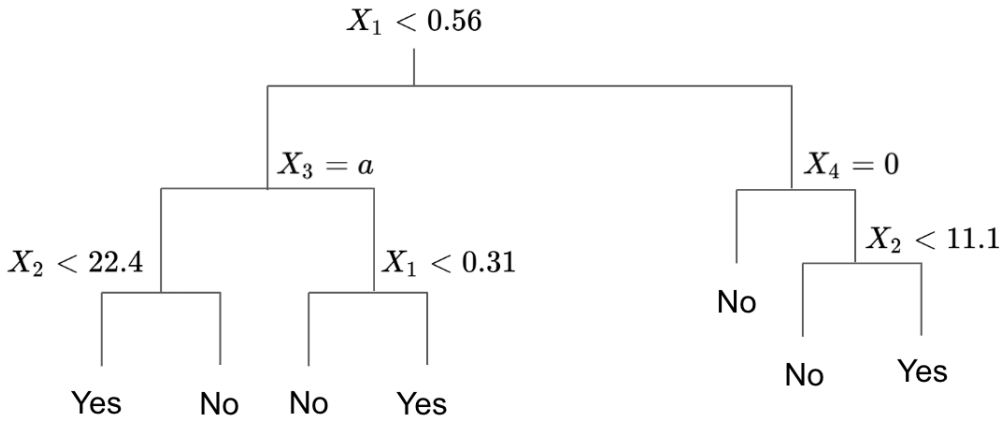


Figure 2:

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

3

Build a decision tree for the following dataset where `Habitable` is the target variable. The algorithm should use the information gain, stop when all instances in the branches have the same class, and you do not need to apply a pruning algorithm.

Size	Orbit	Temperature	Habitable
Big	Far	200	No
Big	Near	200	No
Big	Near	260	Yes
Big	Near	380	Yes
Small	Far	200	Yes
Small	Far	260	Yes

The usual way of dealing with continuous features such as `Temperature` is to sort the values and split according to the midpoint between consecutive values. For the variable `Temperature`, the sorted values would be 200, 260 and 380, and the midpoints would be 230 and 320. This creates two splitting criteria: `Temperature` ≤ 230 and `Temperature` ≤ 320.

As a reminder, to compute the information gain, we need to compute the entropy of Y:

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

and the expected conditional entropy of Y given X:

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

Then, the information gain is given by:

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

Solution:

We denote the features `Size`, `Orbit` and `Temperature` by S, O and T respectively. Furthermore, we denote the target `Habitable` by Y.

We will build our decision tree using the recursive binary splitting algorithm such that, at each split, the information gain of $Y \in \mathcal{Y}$ for a feature $X \in \mathcal{X}$ is maximized. To find out which split leads to the highest information gain, we must account for every variable and for every possible split among 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 usually a minimum number of samples per leaf (e.g. 5). However, for this exercise, we will grow the tree to its full depth.

1. **First iteration**

The entropy of Y is equal to:

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

1.1. Variable `Size`

$$
\begin{aligned}
H(Y|S=\text{Big}) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \\
H(Y|S=\text{Small}) &= 0 \\
H(Y|S) &= p(S=\text{Big})H(Y|S=\text{Big}) + p(S=\text{Small})H(Y|S=\text{Small}) \\
&= \frac{2}{3}\cdot 1 + \frac{1}{3}\cdot 0 \\
&= 0.67
\end{aligned}
$$

1.2. Variable `Orbit`

$$
\begin{aligned}
H(Y|O=\text{Far}) &= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.92 \\
H(Y|O=\text{Near}) &= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.92 \\
H(Y|O) &= p(O=\text{Far})H(Y|O=\text{Far}) + p(O=\text{Near})H(Y|O=\text{Near}) \\
&= \frac{1}{2}\cdot 0.92 + \frac{1}{2}\cdot 0.92 \\
&= 0.92
\end{aligned}
$$

1.3. Variable `Temperature`

From the variable T, we create two boolean random variables: $T_1 = T \leq 230$ and $T_2 = T \leq 320$.

$$
\begin{aligned}
H(Y|T_1=\text{True}) &= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92 \\
H(Y|T_1=\text{False}) &= 0 \\
H(Y|T_1) &= p(T1=\text{True})H(Y|T_1=\text{True}) + p(T_1=\text{False})H(Y|T_1=\text{False}) \\
&= \frac{1}{2}\cdot 0.92 + \frac{1}{2}\cdot 0 \\
&= 0.46
\end{aligned}
$$

$$
\begin{aligned}
H(Y|T_2=\text{True}) &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97 \\
H(Y|T_2=\text{False}) &= 0 \\
H(Y|T_2) &= p(T_2=\text{True})H(Y|T_2=\text{True}) + p(T_2=\text{False})H(Y|T_2=\text{False}) \\
&= \frac{5}{6}\cdot 0.97 + \frac{1}{6}\cdot 0 \\
&= 0.81
\end{aligned}
$$

The feature that gives the maximum information gain is the feature that gives the minimum expected conditional entropy. It is obtained by splitting the training dataset in two regions according to $T \leq 230$. By taking a majority vote on the class of the observations that fall in either of the two regions, we get the decision tree of Figure 3.

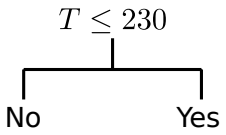


Figure 3: Decision tree obtained after the first step.

As all observations that fall into the region $T > 230$ 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 $T \leq 230$, and so begins the second iteration of the algorithm.

2. **Second iteration**

The region $T \leq 230$ contains the following samples.

Size	Orbit	Temperature	Habitable
Big	Far	200	No
Big	Near	200	No
Small	Far	200	Yes

In the second iteration, all computations are restricted to this dataset. In other words, all probabilities will be conditional to $T \leq 230$. We omit this condition for the sake of brevity.

The entropy of the variable Y in R_1 is given by:

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

2.1. Variable `Size`

$$\begin{aligned} H(Y|S = \text{Big}) &= 0 \\ H(Y|S = \text{Small}) &= 0 \\ H(Y|S) &= p(S = \text{Big})H(Y|S = \text{Big}) + p(S = \text{Small})H(Y|S = \text{Small}) \\ &= \frac{2}{3}\cdot 0 + \frac{1}{3}\cdot 0 \\ &= 0 \end{aligned}$$

2.2. Variable `Orbit`

$$\begin{aligned} H(Y|O = \text{Far}) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \\ H(Y|O = \text{Near}) &= 0 \\ H(Y|O) &= p(O = \text{Far})H(Y|O = \text{Far}) + p(O = \text{Near})H(Y|O = \text{Near}) \\ &= \frac{2}{3}\cdot 1 + \frac{1}{3}\cdot 0 \\ &= 0.67 \end{aligned}$$

2.3. Variable `Temperature`

We can't split further with the variable `Temperature` because all temperatures are the same in this node.

The minimum expected conditional entropy is given by splitting according to `Size`. The resulting decision tree is displayed in Figure 4.

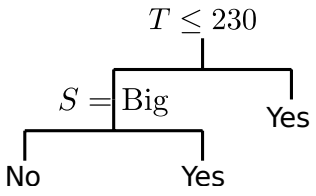


Figure 4: Decision tree obtained after the second step.

All training samples are correctly classified. We cannot further decrease the entropy in any of the regions, thus the algorithm stops.

4

Suppose we produce ten bootstrapped samples from a data set where the label $Y \in \{\text{Green}, \text{Red}\}$ can belong to two classes. We then apply a classification tree to each bootstrapped sample and, for a specific value $X = x$, produce 10 estimates of $p(Y = \text{Red} \mid X = x)$: 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? The classification threshold is set to 0.5, i.e., Red is predicted if $p(Y = \text{Red} \mid X = x) \geq 0.5$.

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

Solution:

Under this threshold, the predicted class for X in each of the bootstrapped samples is: Green, Green, Green, Green, Red, Red, Red, Red, Red, Red. Taking a majority vote, the final predicted class for $X = x$ is Red.

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

5

Prove that, when X and Y are independent random variables, the information gain $\mathrm{IG}(Y|X)$ is null.

Solution:

We know that, when X and Y are independent, $p(y|x) = p(y)$.

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

Thus, $\mathrm{IG}(Y|X) = H(Y) - H(Y|X) = 0$.