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.

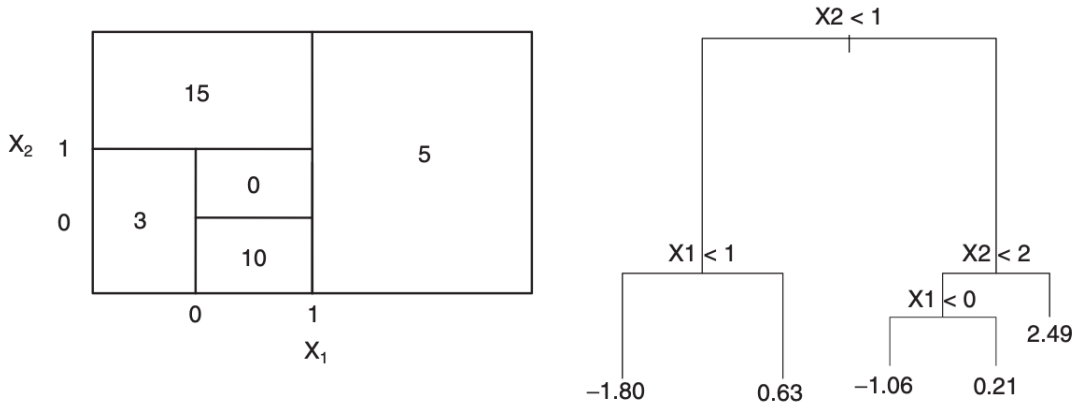

Figure 1:

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

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	

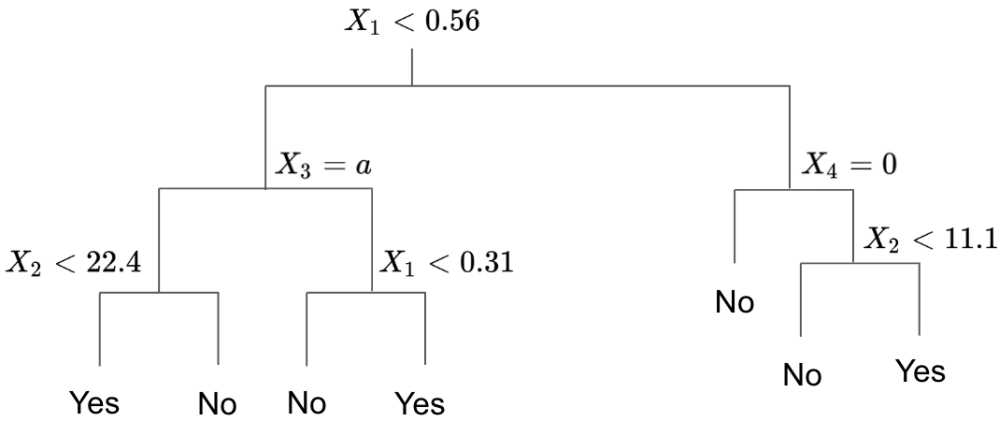

Figure 2:

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:

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

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).

5

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