

COMP5318 Machine Learning and Data Mining

Week 5 Tutorial exercises Decision Trees

Exercise 1. Decision trees and information gain (parts a) and b) – done in class; the rest in your own time)

Consider the following set of training examples:

shape	color	class
circle	blue	+
circle	blue	+
square	blue	-
triangle	blue	-
square	red	+
square	blue	-
square	red	+
circle	red	+

Adapted from M. Kubat, Introduction to Machine Learning, Springer, 2017

- What is the entropy of this collection of training examples with respect to the class?
- What is the information gain of the attribute *shape*?
- Which attribute will be selected as root of the tree based on information gain?
- Build the whole decision tree. Draw the tree after each selected attribute.

You may use this table to calculate information gain:

x	y	-(x/y)* log ₂ (x/y)	x	y	-(x/y)* log ₂ (x/y)	x	y	-(x/y)* log ₂ (x/y)	x	y	-(x/y)* log ₂ (x/y)
1	2	0.50	4	5	0.26	6	7	0.19	5	9	0.47
1	3	0.53	1	6	0.43	1	8	0.38	7	9	0.28
2	3	0.39	5	6	0.22	3	8	0.53	8	9	0.15
1	4	0.5	1	7	0.40	5	8	0.42	1	10	0.33
3	4	0.31	2	7	0.52	7	8	0.17	3	10	0.52
1	5	0.46	3	7	0.52	1	9	0.35	7	10	0.36
2	5	0.53	4	7	0.46	2	9	0.48	9	10	0.14
3	5	0.44	5	7	0.35	4	9	0.52			

Solution:

a) $H(S) = I(5/8, 3/8) = -5/8 \log(5/8) - 3/8 \log(3/8) = 0.42 + 0.53 = 0.95$ bits

b) Split on *shape*:

$$H(S_{\text{circle}}) = I(3/3, 0/3) = -3/3 \log(3/3) - 0/3 \log(0/3) = 0 + 0 = 0 \text{ bits}$$

$$H(S_{\text{square}}) = I(2/4, 2/4) = -2/4 \log(2/4) - 2/4 \log(2/4) = 0.5 + 0.5 = 1 \text{ bit}$$

$$H(S_{\text{triangle}}) = I(1/1, 0/1) = -1/1 \log(1/1) - 0/1 \log(0/1) = 0 + 0 = 0 \text{ bits}$$

$$H(S|\text{shape}) = 3/8 * 0 + 4/8 * 1 + 1/8 * 0 = 0.5 \text{ bits}$$

$$\text{gain}(\text{shape}) = 0.95 - 0.5 = 0.45 \text{ bits}$$

- c) To answer this question we need to calculate the information gain of all attributes. The attribute with the highest information gain will be selected.

There are 2 attributes – *shape* and *color*. We already calculate the information gain for *shape*. Let's do this for *color*.

Split on *color*:

$$H(S_{\text{blue}}) = I(2/5, 3/5) = -2/5 \log(2/5) - 3/5 \log(3/5) = 0.53 + 0.44 = 0.97 \text{ bits}$$

$$H(S_{\text{red}}) = I(3/3, 0/3) = -3/3 \log(3/3) - 0/3 \log(0/3) = 0 + 0 = 1 \text{ bit}$$

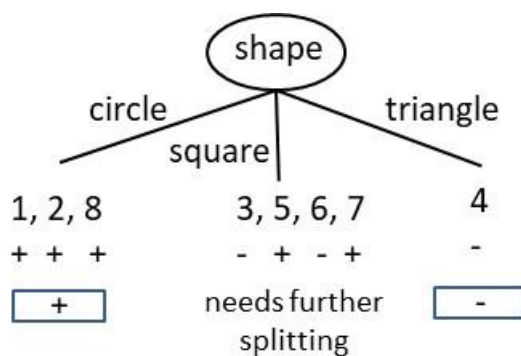
$$H(S|\text{color}) = 5/8 * 0.97 + 3/8 * 0 = 0.61 \text{ bits}$$

$$\text{gain}(\text{color}) = 0.95 - 0.61 = 0.34 \text{ bits}$$

$\text{gain}(\text{shape}) > \text{gain}(\text{color}) \Rightarrow$ *shape* will be selected as the root of the DT (the first attribute to split on)

- d) Building the decision tree:

After selecting *shape*:



We need to repeat the procedure for the examples in the middle branch. The final decision tree is:

