Exercises Trees

1 Decision Trees: Binary Classification

Consider the following data with features A,B,C and the binary class *target*.

A	B	C	target
T	T	1.0	+
T	T	6.0	+
T	F	5.0	-
F	F	4.0	+
F	T	7.0	-
F	T	3.0	-
F	F	8.0	-
T	F	7.0	+
F	T	5.0	-

Gini

Split A = T:

$G(D, (A, A=T)) = \frac{4}{9} \cdot h(\frac{3}{4}) + \frac{5}{9} \cdot h(\frac{1}{5}) = 0{,}3444$

Split B = T:

$G(D, (B, B=T)) = \frac{5}{9} \cdot h(\frac{2}{5}) + \frac{4}{9} \cdot h(\frac{2}{4}) = 0{,}4167$

Split C:

$G(D, (C, 2)) = \frac{1}{9} \cdot h(1) + \frac{8}{9} \cdot h(\frac{3}{8}) = 0{,}4167$

$G(D, (C, 3{,}5)) = \frac{2}{9} \cdot h(\frac{2}{2}) + \frac{7}{9} \cdot h(\frac{3}{7}) = 0{,}4921$

$G(D, (C, 4{,}5)) = \frac{3}{9} \cdot h(\frac{2}{3}) + \frac{6}{9} \cdot h(\frac{2}{6}) = 0{,}4444$

$G(D, (C, 5{,}5)) = \frac{5}{9} \cdot h(\frac{2}{5}) + \frac{4}{9} \cdot h(\frac{2}{4}) = 0{,}4889$

$G(D, (C, 6{,}5)) = \frac{3}{9} \cdot h(\frac{1}{3}) + \frac{6}{9} \cdot h(\frac{3}{6}) = 0{,}4815$

$G(D, (C, 7{,}5)) = \frac{1}{9} \cdot h(0) + \frac{8}{9} \cdot h(\frac{4}{8}) = 0{,}4444$

Missclassification

= 0,2222

= 0,4444

= 0,3333

= 0,4444

= 0,3333

= 0,4444

= 0,4444

= 0,4444

Entropy

Split A = T:

$G(D, (A, A=T)) = \frac{4}{9} \cdot h(\frac{3}{4}) + \frac{5}{9} \cdot h(\frac{1}{5}) = 0{,}762$

Split B = T:

$G(D, (B, B=T)) = \frac{5}{9} \cdot h(\frac{2}{5}) + \frac{4}{9} \cdot h(\frac{2}{4}) = 0{,}984$

Split C:

$G(D, (C, 2)) = \frac{1}{9} \cdot h(1) + \frac{8}{9} \cdot h(\frac{3}{8}) = 0{,}8484$

$G(D, (C, 3{,}5)) = \frac{2}{9} \cdot h(\frac{2}{2}) + \frac{7}{9} \cdot h(\frac{3}{7}) = 0{,}9646$

$G(D, (C, 4{,}5)) = \frac{3}{9} \cdot h(\frac{2}{3}) + \frac{6}{9} \cdot h(\frac{2}{6}) = 0{,}9183$

$G(D, (C, 5{,}5)) = \frac{5}{9} \cdot h(\frac{2}{5}) + \frac{4}{9} \cdot h(\frac{2}{4}) = 0{,}9839$

$G(D, (C, 6{,}5)) = \frac{3}{9} \cdot h(\frac{1}{3}) + \frac{6}{9} \cdot h(\frac{3}{6}) = 0{,}9728$

$G(D, (C, 7{,}5)) = \frac{1}{9} \cdot h(0) + \frac{8}{9} \cdot h(\frac{4}{8}) = 0{,}8889$

Consider all three impurity measures: What is the best split according to each of the three measures?

2 Decision Trees: Binary Classification (XOR)

Consider the following data. Note that the target column is obtained by $A + B + C \mod 2$ (in Boolean logic this is called eXclusive OR - XOR -).

A	B	C	target
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

1. Draw the decision tree (class is *target*) obtained by tree induction using Gini impurity (if the gain of two features is equal, you are allowed to choose one of them randomly).
2. Does the tree change if you use entropy?
3. Suppose you use **prepruning** (so you stop growing the tree) by requiring more than one instance per leaf. How does your tree look like?