

# Homework Set 1, 2020

Cursus: BETA-INFOMDM Data Mining (INFOMDM)

---

**Aantal vragen:** 5

# Homework Set 1, 2020

Cursus: Data Mining (INFOMDM)

---

This is homework set 1 of Data Mining

**Aantal vragen:** 5

- 1 We want to determine the best split in a node containing the following data on numeric attribute x and class label y. The class label can take on three different values, coded as A, B and C.

x	2	2	3	4	4	5	6	7	8	9
y	A	A	A	B	B	B	A	C	C	C

We use the gini-index as impurity measure.

What is the best split on x?

- a. Between x=2 and x=3
- b. Between x=3 and x=4
- c. Between x=4 and x=5
- d. Between x=5 and x=6
- e. Between x=6 and x=7**
- f. Between x=7 and x=8
- g. Between x=8 and x=9

- 2 We want to determine the optimal split in a node that contains the following data:

x	a	a	a	b	b	c	c	d	d
y	0	0	1	0	1	1	1	0	0

Here x is a categorical attribute with possible values {a,b,c,d} and y is a binary class label with values 0 and 1. We use the gini-index as impurity measure.

The best split on x is:

- a.  $x \in \{a,b\}$
- b.  $x \in \{a,c\}$
- c.  $x \in \{a,b,c\}$**
- d.  $x \in \{a,b,d\}$
- e.  $x \in \{a,d\}$
- f.  $x \in \{b,c,d\}$
- g.  $x \in \{a,c,d\}$

- 3 The following binary classification tree has been grown on a training set with  $n=100$  examples:

Node	t1	t2	t3 - leaf	t4 - leaf	t5	t6	t7 - leaf	t8 - leaf	t9 - leaf
Class A	50	30	30	0	20	10	10	0	10
Class B	50	10	0	10	40	40	0	40	0

The nodes have been numbered according to depth first preorder traversal. We listed the number of cases a node contains of each class below it.

Perform cost-complexity pruning on this tree, and answer the following questions.

The value of  $\alpha_2 =$  **a.**  $\frac{1}{10}$  ()

The value of  $\alpha_3 =$  **b.**  $\frac{2}{10}$  ()

- 4 Let  $n$  denote the number of observations and  $k$  the number of classes. Let  $x$  be a numeric variable with all values distinct. We use the gini-index as impurity measure. Let  $S$  denote the number of splits we have to evaluate in order to determine the best split on  $x$ .

Give an expression for  $S$  in the most unfavorable case:  $S =$  **a.** ..... ()  $(n-1)((k-1)/k)$

Give an expression for  $S$  in the most favorable case:  $S =$  **b.** ..... ()  $k-1$

**5** In learning classification trees, determination of the appropriate size of the tree is an important problem. One can control the size of the tree by using a so-called stopping rule to stop growing the tree early. One possibility to implement this idea is to use parameters  $n_{min}$  and  $minleaf$ . If a node contains less than  $n_{min}$  cases, then it becomes a leaf node. A split is not allowed if it creates a child node with less than  $minleaf$  cases. Assume the tree growing algorithm only makes binary splits.

**a.** Consider the following two parameter settings:

1.  $n_{min}=12$  and  $minleaf=10$

2.  $n_{min}=18$  and  $minleaf=10$

Would you expect the tree in case (1) to have a lower, higher, or the same error rate on the training sample (resubstitution error) as the tree in case (2)?

☒ **a.** lower

☐ **b.** higher

☐ **c.** the same

**b.** Answer the same question for the following parameter settings:

1.  $n_{min}=20$  and  $minleaf=5$

2.  $n_{min}=10$  and  $minleaf=5$

☐ **a.** lower

☒ **b.** higher

☐ **c.** the same

Thank you, goodbye!