

# Data Mining 2020

## Exercises Classification Trees

### Exercise 1: Computing Splits

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

$x_1$	c	b	b	a	a	b	e	e	d	e
$x_2$	28	31	35	40	40	45	45	52	52	60
$y$	B	B	B	A	B	A	B	A	A	A

Here  $x_1$  is a categorical attribute with possible values  $\{a,b,c,d,e\}$ ,  $x_2$  is a numerical attribute, and  $y$  is a binary class label with possible values A and B. We use the gini-index as impurity measure. The best split is the one that maximizes the impurity reduction.

- (a) How many possible binary splits are there on  $x_1$ ?
- (b) How many splits on  $x_1$  do we have to evaluate to determine the best one? List them.
- (c) How many possible binary splits are there on  $x_2$ ?
- (d) How many splits on  $x_2$  do we have to evaluate to determine the best one? List them. (Use the fact that the best split can not occur inside a segment.)
- (e) Give the impurity reduction of the best split on  $x_2$ .

### Exercise 2: More On Computing Splits

Consider 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.

$x$	6	8	12	12	12	14	14	14	18	20
$y$	A	A	A	A	B	A	A	B	C	C

We use the gini-index as impurity measure. The formula for the gini-index for an arbitrary number of class labels is given by

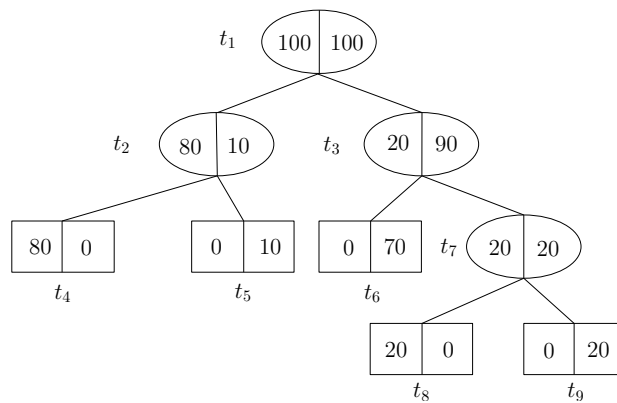
$$i(t) = 1 - \sum_{j=1}^c p(j|t)^2,$$

where  $C$  denotes the number of class labels, and  $p(j|t)$  denotes the relative frequency of class  $j$  in node  $t$ .

- Which candidate split(s) do we have to evaluate to determine the best one? (don't list any more than strictly necessary)
- What is the best split on  $x$ , and what is the impurity reduction of that split?
- Suppose we have the constraint  $\text{minleaf}=3$ , that is, you are not allowed to create a child node with less than 3 data points. Give the best split on  $x$  that satisfies the  $\text{minleaf}$  constraint. Is it on the border of a segment?

### Exercise 3: Cost-Complexity Pruning

The tree  $T_{\max}$  given below has been grown on the training sample.



In each node the number of observations with class A is given in the left part, and the number of observations with class B in the right part. The leaf nodes have been drawn as rectangles. The total cost of a tree  $T$  is defined as:

$$C_{\alpha}(T) = R(T) + \alpha|\tilde{T}| \quad (1)$$

It can be written as the sum of the contribution of each leaf node to total cost:

$$C_{\alpha}(T) = \sum_{t \in \tilde{T}} (R(t) + \alpha), \quad (2)$$

where  $R(t)$  is the number of classification errors made in node  $t$ , divided by the total number of observations in the training set. For  $T_{\max}$  as given above, this is:

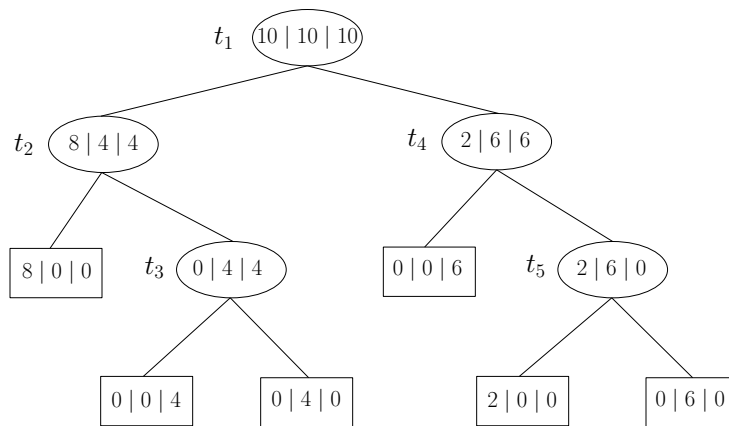
$$C_{\alpha}(T_{\max}) = (R(t_4) + \alpha) + (R(t_5) + \alpha) + (R(t_6) + \alpha) + (R(t_8) + \alpha) + (R(t_9) + \alpha) \quad (3)$$

- As was done for  $T_{\max}$  in equation (3), give an expression for the total cost of  $T_{\max} - T_{t_3}$ , the tree obtained by pruning  $T_{\max}$  in  $t_3$ .

- (b) Which terms are present in the expression for total cost of  $T_{\max}$  but not  $T_{\max} - T_{t_3}$ ?  
Which terms are present in the expression for total cost of  $T_{\max} - T_{t_3}$  but not  $T_{\max}$ ?
- (c) For what value of  $\alpha$  is the total cost of  $T_{\max}$  and  $T_{\max} - T_{t_3}$  the same?  
Which tree is preferred in that case?
- (d) Give  $T_1 = T(\alpha = 0)$ : the smallest minimizing subtree of  $T_{\max}$  for  $\alpha = 0$ .
- (e) Compute the cost-complexity sequence  $T_1 > T_2 > \dots > \{t_1\}$ .  
Also give the corresponding sequence of  $\alpha$  values.

### Exercise 4: Cost-Complexity Pruning

The tree given below, denoted by  $T_{\max}$ , has been constructed on the training sample:



In each node, the number of observations with class A is given in the left part, the number of observations with class B is given in the middle part, and the number of observations with class C is given in the right part. The leaf nodes have been drawn as rectangles.

Compute the cost-complexity pruning sequence  $T_1 > T_2 > \dots > \{t_1\}$ , where  $T_1$  is the smallest minimizing subtree for  $\alpha = 0$ . Also give the corresponding sequence of  $\alpha$  values.

### Exercise 5: An Alternative Pruning Procedure

In their seminal work *Classification and Regression Trees*, Breiman et al. (Chapman & Hall, 1984) consider the following pruning procedure before they describe cost-complexity pruning. Suppose that  $T_{\max}$  has  $L$  terminal nodes. Construct a sequence of smaller and smaller trees

$$T_{\max}, T_1, T_2, \dots, \{t_1\}$$

as follows: For every value of  $H$ ,  $1 \leq H \leq L$ , consider the class  $\mathcal{T}_H$  of all subtrees of  $T_{\max}$  having  $L - H$  leaf nodes. Select  $T_H$  as the subtree in  $\mathcal{T}_H$  which minimizes  $R(T)$ ; that is,

$$R(T_H) = \min_{T \in \mathcal{T}_H} R(T).$$

Put another way,  $T_H$  is the minimal resubstitution error pruned subtree of  $T_{\max}$  having  $L - H$  leaf nodes.

- (a) Give the sequence

$$T_{\max}, T_1, T_2, \dots, \{t_1\}$$

obtained when you apply this pruning method to the tree  $T_{\max}$  given in exercise 3.

- (b) Does the sequence you obtained under (a) have the desirable property that the sequence is nested, i.e., do we have

$$T_{\max} > T_1 > T_2 > \dots > \{t_1\}?$$

- (c) Is the sequence of minimal cost-complexity trees a subsequence of the sequence of subtrees as defined above? In other words, if  $T(\alpha)$  has  $m$  leaf nodes, can there be another subtree  $T$  having  $m$  leaf nodes with  $R(T) \leq R(T(\alpha))$ ?

## Exercise 6: The Gini index

We have defined the gini index for binary classification as

$$i(t) = p(0|t)p(1|t) = p(0|t)(1 - p(0|t)), \quad (4)$$

where the class values are coded as 0 and 1, and  $p(j|t)$  denotes the relative frequency of class  $j$  in node  $t$ . The generalization to an arbitrary number of classes is given by:

$$i(t) = \sum_{j=1}^C p(j|t)(1 - p(j|t)), \quad (5)$$

where  $C$  denotes the number of classes.

- (a) If we apply equation (5) to the binary case, we should get the same results as when we apply equation (4). Is this indeed the case?
- (b) Show that equation (5) can alternatively be written as

$$i(t) = 1 - \sum_{j=1}^C p(j|t)^2.$$

## Exercise 7: More about the Gini index

The expected value (mean) of a discrete random variable  $X$  is defined as

$$\mathbb{E}[X] = \sum_x x \times P(X = x),$$

where the sum is over all possible values  $x$  of  $X$ . Furthermore,

$$\mathbb{E}[f(X)] = \sum_x f(x) \times P(X = x).$$

The variance of  $X$  is defined as its expected squared deviation from the mean:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Let  $X \in \{0, 1\}$  be a binary random variable with  $P(X = 1) = p$ , and  $P(X = 0) = 1 - p$ . We also say that  $X$  has a Bernoulli distribution. Show that:

- (a)  $\mathbb{E}[X] = p$ , and
- (b)  $\mathbb{V}[X] = p(1 - p)$ .
- (c) Gini impurity is defined as  $\phi(p) = p(1 - p)$  for  $0 \leq p \leq 1$ . Use calculus to show that this function achieves its maximum for  $p = \frac{1}{2}$ .
- (d) Use calculus to show that the Gini index is strictly concave.

## Exercise 8: The binomial distribution

Let  $Y$  denote the fraction of *ones* in  $n$  independent Bernoulli trials:

$$Y = \frac{1}{n} \sum_{i=1}^n X_i$$

where  $X_1, \dots, X_n$  are independent Bernoulli random variables, with  $P(X_i = 1) = p$ , for  $i = 1, \dots, n$ .

- (a) Show that

$$\mathbb{E}[Y] = p.$$

- (b) Show that

$$\mathbb{V}[Y] = \frac{p(1 - p)}{n}.$$

**Turn page over for rules of expectation and variance!**

### Some Useful Properties of Expectation and Variance

1.  $\mathbb{E}(c) = c$  for constant  $c$ . “The expected value of a constant is the constant itself”.
2.  $\mathbb{E}(cX) = c\mathbb{E}(X)$ .
3.  $\mathbb{E}(X \pm Y) = \mathbb{E}(X) \pm \mathbb{E}(Y)$ .
4.  $\mathbb{V}(c) = 0$  for constant  $c$ . “The variance of a constant is zero”.
5.  $\mathbb{V}(cX) = c^2\mathbb{V}(X)$ . “The variance of a constant times a random variable is equal to the square of the constant times the variance of the random variable”.
6.  $\mathbb{V}(X \pm Y) = \mathbb{V}(X) + \mathbb{V}(Y)$  if  $X$  and  $Y$  are independent.

More generally, let  $Z = c_0 + \sum_{i=1}^n c_i X_i$ . Then

1.  $\mathbb{E}(Z) = \mathbb{E}(c_0 + \sum_{i=1}^n c_i X_i) = c_0 + \sum_{i=1}^n c_i \mathbb{E}(X_i)$
2.  $\mathbb{V}(Z) = \mathbb{V}(c_0 + \sum_{i=1}^n c_i X_i) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i)$ , provided that the  $X_i$  are mutually independent.