Data Mining: Exercises on Classification Trees

Exercise 1: Computing Splits

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

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

Here x_1 is a categorical attribute with possible values {a,b,c,d,e}, x_2 is a numerical attribute, and y is the binary class label. We use the gini index as splitting criterion.

(a) How many possible binary splits are there on x_1?

(b) Which splits on x_1 do we have to evaluate to determine the optimal one?

(c) How many possible binary splits are there on x_2?

(d) Which splits on x_2 do we have to evaluate to determine the optimal one?

Exercise 2: Splitting and Pruning

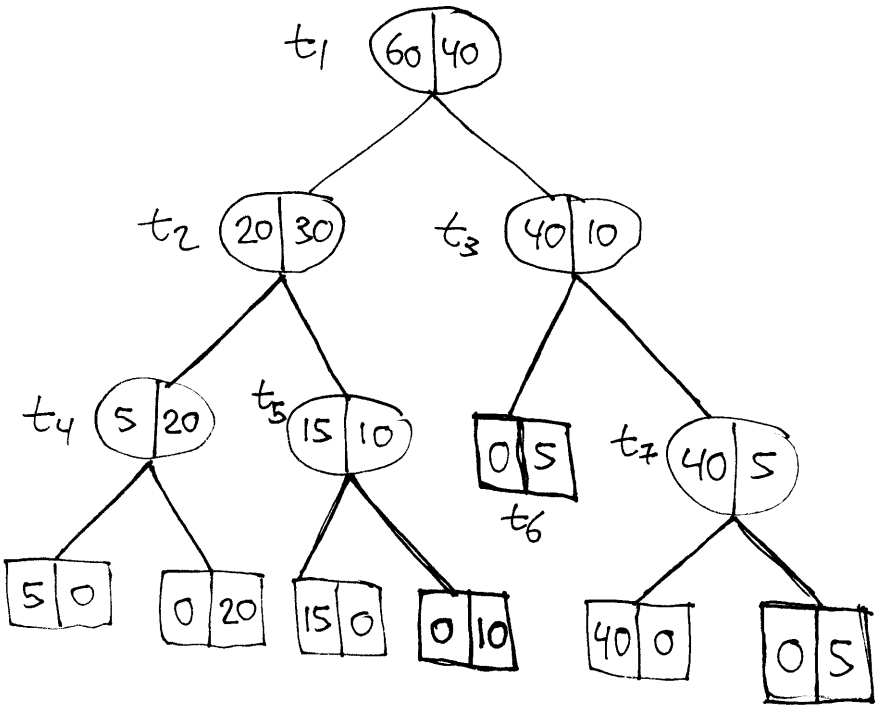
We have grown the tree at the top of the next page (called $T_{\max}$) on a training sample. In each node, the number of observations with class A is given in the left part, and the number of observations with class B in the right part. The leaf nodes have been drawn as rectangles (sort of).

(a) Compute the impurity of nodes t_1, t_2 and t_3 using the gini index.

(b) Give the impurity reduction achieved by the first split.

(c) Compute T_1, the smallest minimizing subtree of $T_{\max}$ for $\alpha = 0$.

(d) Compute the cost-complexity pruning sequence $T_1 > T_2 > \ldots > \{t_1\}$. For each tree in the sequence, give the interval of α values for which it is the smallest minimizing subtree of $T_{\max}$.



Exercise 3: Gini index

We have defined the gini index for binary classification as

$$i(t) = p(0|t)p(1|t) = p(0|t)(1 - p(0|t)), \tag{1}$$

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:

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

where C denotes the number of classes. If we apply equation (2) to the binary case, we should get the same results as when we apply equation (1). Is this indeed the case?

Show that equation (2) can alternatively be written as

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