Exercise 8.1 CART (Part 1) Solutions

Understanding the Search for Best Split:

1. In default10.xlsx, the training set consists of 10 cases with 3 input variables (Home Owner, Marital Status, and Annual Income). The outcome variable is Defaulted Borrower. Let y= 0 for no default, and y = 1 for defaulted. Using Excel for all calculations,

   (a) What are the proportions of cases before any splitting?

   (b) Calculate the Entropy, Gini index and Misclassification error of the two descendant nodes if the splitting criteria is:

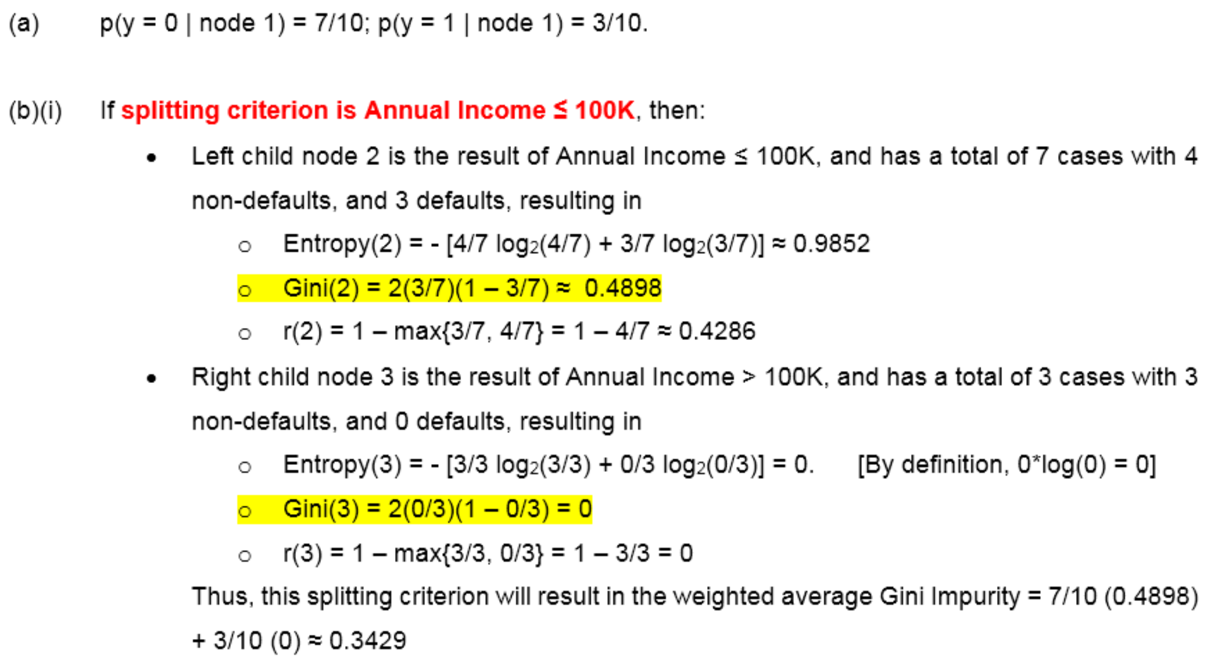
   (i) Annual Income $\leqslant$ 100K

   (ii) Marital Status = Single

   (c) Breiman et al (1984) preferred the Gini index. Which of the above 2 split would you choose if the Gini index is used?

   (d) Misclassification error is not a good formula for selecting the best split, in general. Why?

Note: Here, we had only compared 2 possible splits. CART will have to consider and compare all possible splits from all the X variables and variable values in order to find the best split.



(a) p(y = 0 | node 1) = 7/10; p(y = 1 | node 1) = 3/10.

(b)(i) If **splitting criterion is Annual Income ≤ 100K**, then:

- Left child node 2 is the result of Annual Income ≤ 100K, and has a total of 7 cases with 4 non-defaults, and 3 defaults, resulting in
  - Entropy(2) = - [4/7 $\log_2$(4/7) + 3/7 $\log_2$(3/7)] ≈ 0.9852
  - Gini(2) = 2(3/7)(1 – 3/7) ≈ 0.4898
  - r(2) = 1 – max{3/7, 4/7} = 1 – 4/7 ≈ 0.4286
- Right child node 3 is the result of Annual Income > 100K, and has a total of 3 cases with 3 non-defaults, and 0 defaults, resulting in
  - Entropy(3) = - [3/3 $\log_2$(3/3) + 0/3 $\log_2$(0/3)] = 0. [By definition, 0*log(0) = 0]
  - Gini(3) = 2(0/3)(1 – 0/3) = 0
  - r(3) = 1 – max{3/3, 0/3} = 1 – 3/3 = 0

  Thus, this splitting criterion will result in the weighted average Gini Impurity = 7/10 (0.4898) + 3/10 (0) ≈ 0.3429 (b)(i) If **splitting criterion is Marital Status = Single**, then:

- Left child node 2 is the result of Marital Status = Single, and has a total of 4 cases with 2 non-defaults, and 2 defaults, resulting in
    - Entropy (2) = - [2/4 $\log_2$(2/4) + 2/4 $\log_2$(2/4)] = 1
    - Gini(2) = 2(2/4)(1 – 2/4) = 0.5
    - r(2) = 1 – max{2/4, 2/4} = 1 – 2/4 = 0.5
- Right child node 3 is the result of Marital Status ≠ Single, and has a total of 6 cases with 5 non-defaults, and 1 defaults, resulting in
    - Entropy (3) = - [5/6 $\log_2$(5/6) + 1/6 $\log_2$(1/6)] ≈ 0.65
    - Gini(3) = 2(1/6)(1 – 1/6) ≈ 0.2778
    - r(3) = 1 – max{1/6, 5/6} = 1 – 5/6 = 0.1667

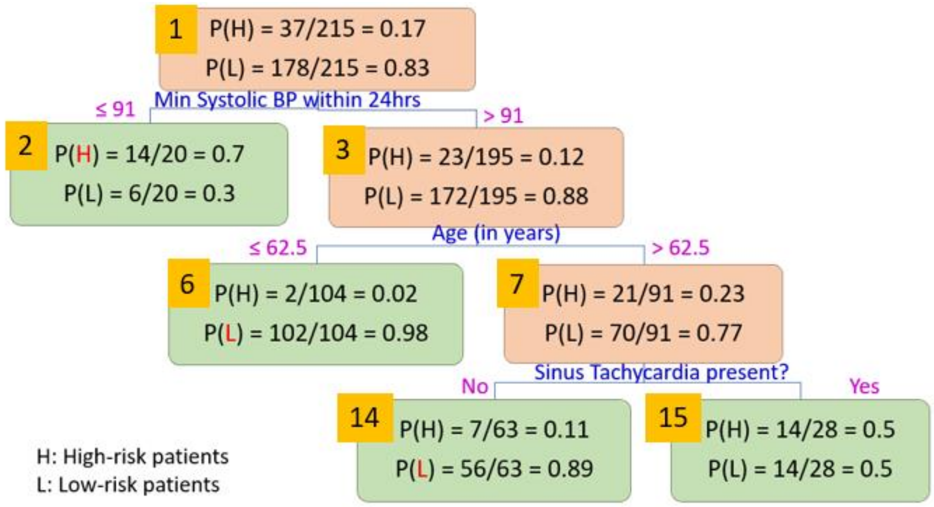
Thus, this splitting criterion will result in the weighted average Gini Impurity = 4/10 (0.5) + 6/10 (0.2778) ≈ 0.3667

(c) Based on minimizing weighted average Gini Impurity, splitting criterion (b)(i) is better than (b)(ii).

Gini improvement = 0.42 – 0.3429 = 0.0771

(d) Misclassification error only considers the majority. i.e. How the errors are distributed among the minorities are ignored. Example: If predicting Y = A, B or C, and P(A) = 0.6, then A is the majority and misclassification error = 0.4. However, how the 40% is distributed among B and C is important too. 20% B and 20% C is “worse” outcome then 1% B, 39% C. Hence, misclassification error is only suitable for use as node impurity measure only for the case where Y has only two categories. Otherwise, it is not suitable.

Understanding the CART Model Structure:



2. What is the decision rule, model prediction and misclassification error at node 6?
If min Systolic BP within 24hrs > 91 and Age ≤ 62.5 years old, then model predict Low risk. r(6) = 0.02.

3. What is the decision rule, model prediction and misclassification error at node 7? N.A. Node 7 is not a terminal node.

An alternative perspective on Pruning by considering the portion of the tree to be pruned away instead of considering the tree that remains after pruning. This perspective directly defines the strength of the link to desendants.

Define T_t to be a subtree with root at node t, and define $R_\alpha(t) = R(t) + \alpha$

If $R_\alpha(T_t) < R_\alpha(t)$, then the contribution to the cost complexity of the subtree T_t is less than that for node t (if it becomes a terminal node). Do not prune.

As α increases, equality is achieved when

$$\alpha^* = \frac{R(t) - R(T_t)}{|T_t| - 1}$$

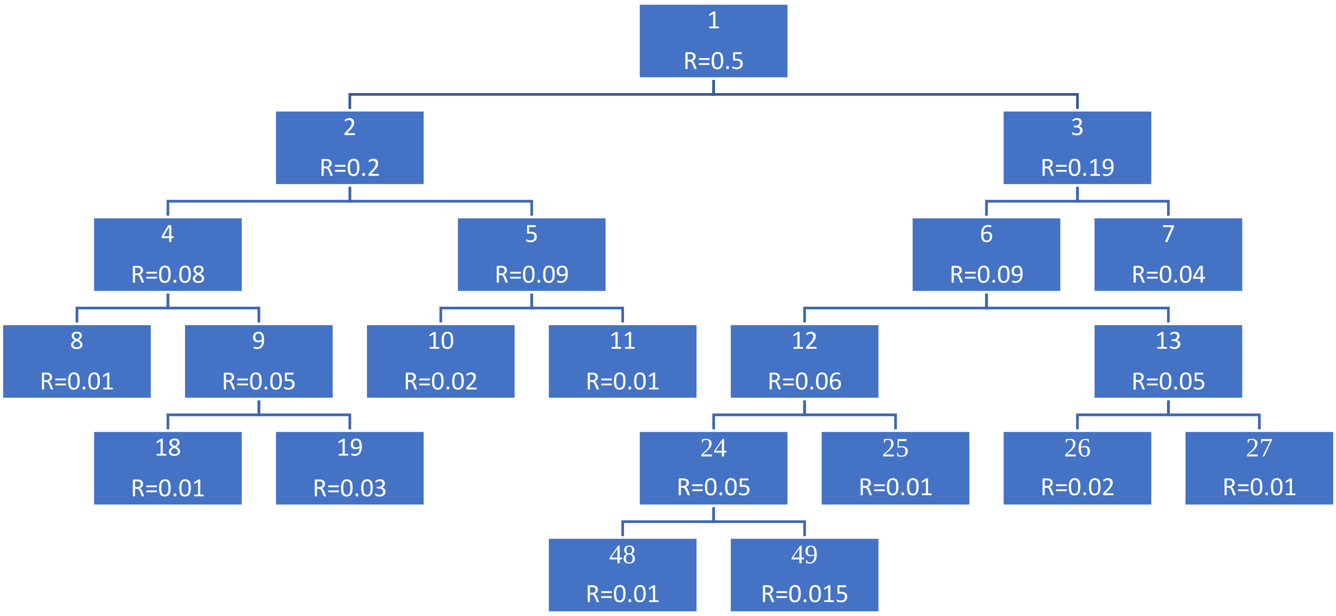
and termination of the tree at t is preferred if $\alpha \geq \alpha^*$ i.e. prune all descendant nodes at node t.

Thus, we can define

$$g(t) = \frac{R(t) - R(T_t)}{|T_t| - 1}$$

as the strength of the link from node t to its descendants.

4. A fully grown tree is shown below with R(t) calculated at each node. Calculate and write down g(t) in the table below to get the weakest link for pruning.



t	$R(t)$	$R(T_t)$	$\|T_t\|-1$	$g(t)$
1	0.5	{R(8)+R(18)+R(19)+R(10)+R(11)} + {R(7)+R(48)+R(49)+R(25)+R(26)+R(27)} = 0.08 + 0.105 = 0.185	10	0.0315
2	0.2	R(8)+R(18)+R(19)+R(10)+R(11) = 0.08	4	0.03
3	0.19	R(7)+R(48)+R(49)+R(25)+R(26))+R(27) = 0.105	5	0.017
4	0.08	R(8)+R(18)+R(19) = 0.05	2	0.015
5	0.09	R(10)+R(11) = 0.03	1	0.06
6	0.09	R(48)+R(49)+R(25)+R(26))+R(27) = 0.065	4	0.00625
7*	0.04			
8*	0.01			
9	0.05	R(18)+R(19) = 0.04	1	0.01
10*	0.02			
11*	0.01			
12	0.06	R(48)+R(49)+R(25) = 0.035	2	0.0125
13	0.05	R(26))+R(27) = 0.03	1	0.02
18*	0.01			
19*	0.03			
24	0.05	R(48)+R(49) = 0.025	1	0.025
25*	0.01			
26*	0.02			
27*	0.01			
48*	0.01			
49*	0.015			

Note: * is used to label terminal nodes in this table.

The CART Pruning Sequence

After growing the tree using the Gini impurity index, and obtaining g(t) for all internal nodes and R(t) for all nodes. The CART pruning sequence is as follows:

1. Find the node t with the lowest g(t). i.e. node t_X where $g(t_X) \leq g(t_i)$, for all nodes $t_i \neq t_X$
   (There maybe more than one node with the minimum g(t) value.)

2. Prune at nodes t_X.
3. Re-compute g(t) at all ancestor (and internal) nodes of t_X
4. Repeat steps 1 to 3 until only the root node remains.