RPART Kyphosis "Improvement" Output STATS 784

For the root, "improve = 6.762330 (0 missing)" for "Start < 8.5".

Q: What is this?

A: According to techrept.ps, $I(A) = \sum_{i=1}^{c} f(p_{iA})$ where p_{iA} = proportion of those in A belonging to class i for future samples,

info index $f(p) = -p \log p$ and

Gini index $f(p) = p(1-p)$,

A is some node of the tree.

Then $\Delta I = p(A)\, I(A) - \left[p(A_L)\, I(A_L) + p(A_R)\, I(A_R)\right]$.

For Kyphosis there are 2 classes. Also, Gini is used by default for method = "class".

Therefore $I(A) = p(1-p) + (1-p)p = 2p(1-p)$.

At the root

presence	absence	
17	64	81

Also, for nodes 2 & 3,

	presence	absence	
Start < 8.5	6	56	62
Start ≥ 8.5	11	8	19

$$\Delta I(\text{root}) = 81 \times 2 \times \frac{17}{81} \times \frac{64}{81} - 62 \times 2 \times \frac{6}{62} \times \frac{56}{62} -$$

$$19 \times 2 \times \frac{11}{19} \times \frac{8}{19}$$

$$\doteq 6.762330.$$

Note that S-PLUS has "Start < 12.5" at root. This results in

	presence	absence
Start ≤ 12.5	15	20
Start ≥ 12.5	2	44

Improvement (according to rpart()) is

$$\Delta I(\text{root}) = 81 \times 2 \times \frac{17}{81} \times \frac{64}{81} - 35 \times 2 \times \frac{15}{35} \times \frac{20}{35} - 46 \times 2 \times \frac{2}{46} \times \frac{44}{46}$$

$$= 5.8953$$

This is less than 6.762330, which is to be expected.