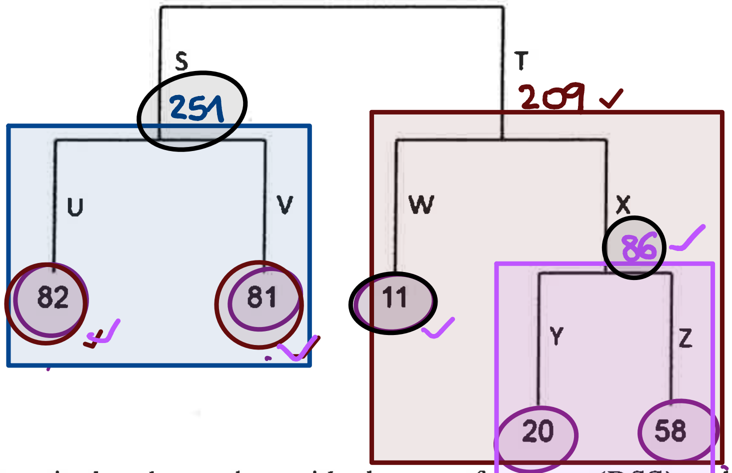
38:

You are given the following unpruned decision tree:



The values at each terminal node are the residual sums of squares (RSS) at that node. The table below gives the RSS at nodes S, T, and X if the tree was pruned at those nodes:

Node	RSS
S	251
T	209
X	86

The RSS for the null model is 486. You use the cost complexity pruning algorithm with the tuning parameter, α, equal to 9 in order to evaluate the following pruning strategies.

I. No nodes pruned $82+81+11+20+58+9\cdot 5=297$

II. Prune node S only $251+11+20+58+9\cdot 4=376$

III. Prune node T only $82+81+209+9\cdot 3=399$

IV. Prune node X only $82+81+11+86+9\cdot 4=296$

V. Prune both nodes S and X $251+11+86+9\cdot 3=375$

Determine which pruning strategy is selected.

✗ A. I

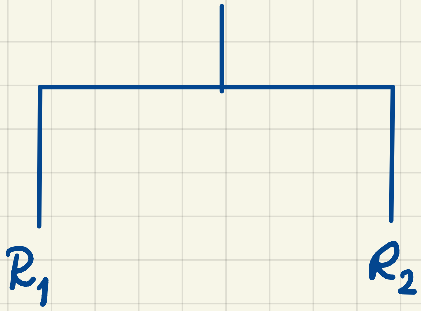
✗ B. II

✗ C. III

[✓ D. IV]

✗ E. V

CONTINUED ON NEXT PAGE

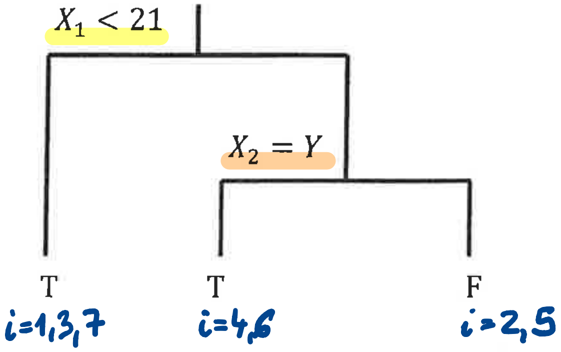


$$\sum_{i\in R_1}(y_i-\hat{y}_1)^2 \qquad \sum_{i\in R_2}(y_i-\hat{y}_2)^2$$

w/ $\hat{y}_1 = \frac{1}{|R_1|}\sum_{i\in R_1} y_i$ w/ $\hat{y}_2 = \frac{1}{|R_2|}\sum_{i\in R_2} y_i$

40.

You are given the following classification decision tree and data set:



i	X_1	X_2	Y
1	12	Y	T
2	23	N	F
3	4	Y	F
4	32	Y	F
5	22	N	T
6	30	Y	T
7	18	N	T

Determine the relationship between the classification error rate, the Gini index, and the cross-entropy, summed across all nodes.

A. cross-entropy > Gini index > classification error rate
B. cross-entropy > Gini index = classification error rate
C. classification error rate > Gini index > cross-entropy
D. Gini index > cross-entropy > classification error rate
E. The answer is not given by (A), (B), (C), or (D).

CONTINUED ON NEXT PAGE

→: Caveat: They explicitly say "summed across all nodes" which is different from computing a weighted average.

For $X_1 < 21$, we have observations: $i = 1, 3, 7$ and they have $Y_1 = T$, $Y_2 = F$, $Y_3 = T$

$\Rightarrow$ We would classify any observation w/ $X_1 < 21$ as T

$\Rightarrow$ The classification error is $\boxed{\frac{1}{3}}$

For $X_2 \geq 21$, i.e., all the other observations,

we have that • for $X_2 = Y$, only $i = 4, 6$ would be @ that terminal node

$\Rightarrow$ the classification error is $\boxed{\frac{1}{2}}$

• for $X_2 = N$, $i = 2, 5$ are @ that terminal node

$\Rightarrow$ The classification error is $\boxed{\frac{1}{2}}$.

Total classification error: $\frac{1}{3} + \frac{1}{2} + \frac{1}{2} = \frac{4}{3} = \frac{12}{9}$ ✓ $= 1.33$

At the first node, the Gini index is $(\frac{1}{3})(\frac{2}{3}) + (\frac{2}{3})(\frac{1}{3}) = \frac{4}{9}$

At the second node, the Gini index is $\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$

Athe the third node, the same.

$\Rightarrow$ Total Gini index $\frac{1}{2} + \frac{1}{2} + \frac{4}{9} = \frac{13}{9} = 1.44$

The cross entropy @ first node: $-\frac{1}{3} \ln(\frac{1}{3}) - \frac{2}{3} \ln(\frac{2}{3})$

The cross entropy @ second and third nodes:

$-\frac{1}{2} \ln(\frac{1}{2}) - \frac{1}{2} \ln(\frac{1}{2})$

The sum is: 2.022809