Pessimistic Error Pruning (PEP)

Pessimistic pruning uses the training set to estimate error rates.
An estimate of accuracy or error based on the training set is overly optimistic and, therefore, strongly biased.
The pessimistic pruning method therefore adjusts the error rates obtained from the training set by adding a penalty, so as to counter the bias incurred.

Using **training** set to estimate error on new data:

Error: $\dfrac{N-Nc}{N}$

Adding a penalty:

Pessimistic Error: $\dfrac{N-Nc+0.5}{N}$

N - Number of examples
Nc - Number of examples of the majority class

Error of a node v (if pruned)

$$q(v) = \frac{Nv-Nc,v+0.5}{Nv}$$

where

Nv = Number examples at node v

Nc,v = Number examples in **majority** class at node v

Error of a subtree T

$$q(T) = \frac{\sum\limits_{l\in leafs(T)} (N_l - N_{c,l} + 0.5)}{\sum\limits_{l\in leafs(T)} N_l}$$

Where
l = leaf node of sub-tree T

Prune if

$$q(v) \leq q(T)$$

$Q(V) = (32-20+0.5)/32 = 0.39$

$Q(T) = (20-16+0.5+12-8+0.5)/(12+20) = 9/32 = 0.28$

34	18

20	12

14	6

16	4

4	8

Class = Yes	20
Class = No	10
Error = 10/30	

Training Error (Before splitting) = 10/30

Pessimistic error = (10 + 0.5)/30 = 10.5/30

Training Error (After splitting) = 9/30

Pessimistic error (After splitting)

= (9 + 4 × 0.5)/30 = 11/30

PRUNE!

A?

A1   A2   A3   A4

Class = Yes	8
Class = No	4

Class = Yes	3
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1