C4.5 - pruning decision trees

Quiz 1

Universiteit Leiden

Quiz 1

Q: Is a tree with only pure leafs always the best classifier you can have?

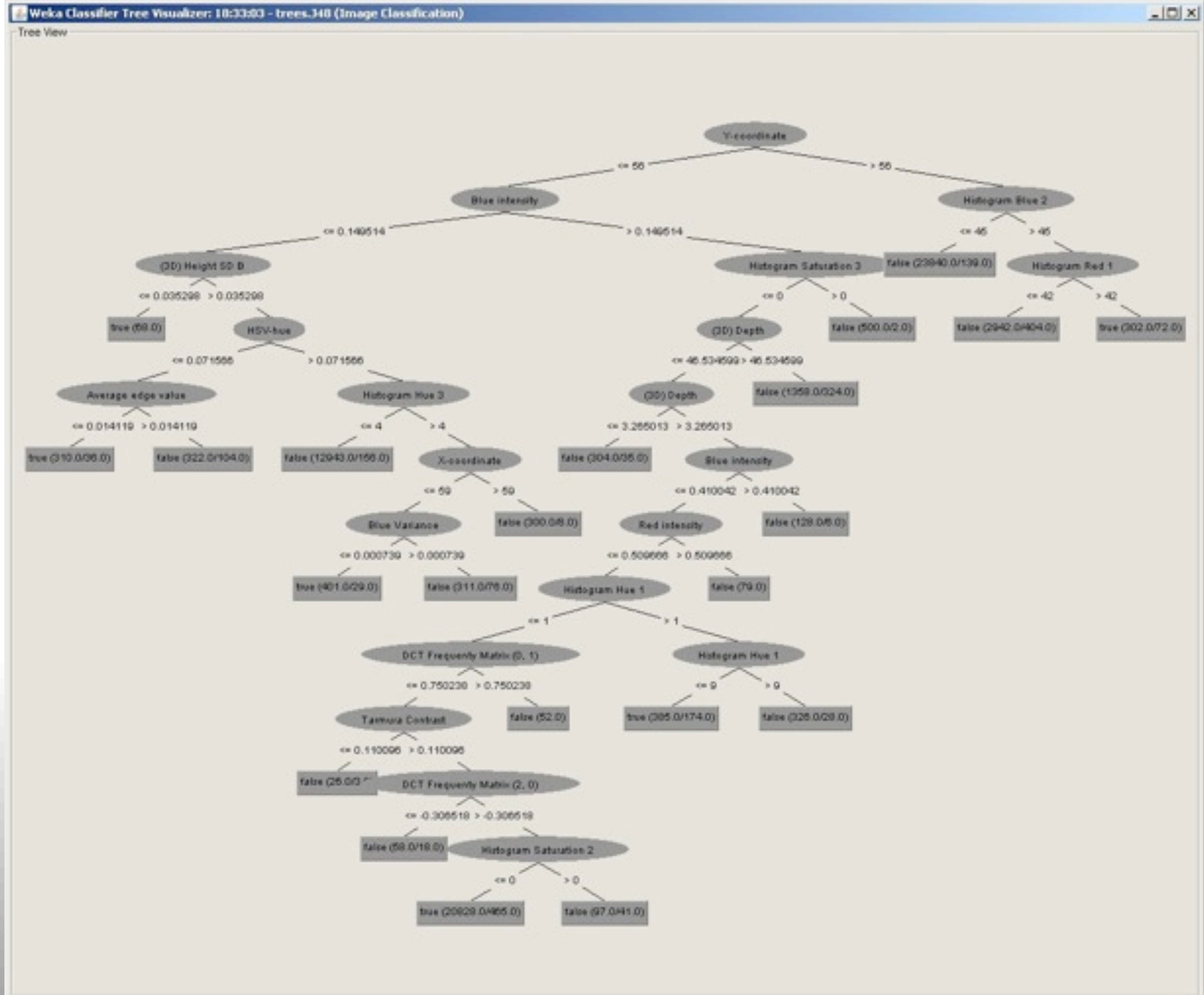
A: No.

Quiz 1

Q: Is a tree with only pure leafs always the best classifier you can have?

A: No.

This tree is the best classifier on the training set, but possibly not on new and unseen data. Because of overfitting, the tree may not generalize very well.


Weka Classifier Tree Visualizer: 18:33:03 - trees.J48 (Image Classification)
Tree View
Y-coordinate
<= 56
> 56
Blue intensity
Histogram Blue 2
<= 0.149514
> 0.149514
<= 46
> 46
(3D) Height SD B
Histogram Saturation 3
false (23940.0/139.0)
Histogram Red 1
<= 0.035298
> 0.035298
<= 0
> 0
<= 42
> 42
true (59.0)
HSV-hue
(3D) Depth
false (500.0/2.0)
false (2942.0/404.0)
true (302.0/72.0)
<= 0.071566
> 0.071566
<= 46.534599
> 46.534599
Average edge value
Histogram Hue 3
(3D) Depth
false (1359.0/324.0)
<= 0.014119
> 0.014119
<= 4
> 4
<= 3.265013
> 3.265013
true (310.0/36.0)
false (322.0/104.0)
false (12943.0/156.0)
X-coordinate
false (304.0/35.0)
Blue intensity
<= 59
> 59
<= 0.410042
> 0.410042
Blue Variance
false (300.0/8.0)
Red intensity
false (128.0/6.0)
<= 0.000739
> 0.000739
<= 0.509666
> 0.509666
true (401.0/29.0)
false (311.0/76.0)
Histogram Hue 1
false (79.0)
<= 1
> 1
DCT Frequency Matrix (0, 1)
Histogram Hue 1
<= 0.750238
> 0.750238
<= 9
> 9
Tamura Contrast
false (52.0)
true (385.0/174.0)
false (326.0/28.0)
<= 0.110096
> 0.110096
DCT Frequency Matrix (2, 0)
<= -0.306518
> -0.306518
false (58.0/18.0)
Histogram Saturation 2
<= 0
> 0
true (20828.0/465.0)
false (97.0/41.0)

Pruning

- Goal: Prevent overfitting to noise in the data
- Two strategies for “pruning” the decision tree:
  - Postpruning - take a fully-grown decision tree and discard unreliable parts
  - Prepruning - stop growing a branch when information becomes unreliable

Prepruning

- Based on statistical significance test
  - Stop growing the tree when there is no statistically significant association between any attribute and the class at a particular node
- Most popular test: chi-squared test
- ID3 used chi-squared test in addition to information gain
  - Only statistically significant attributes were allowed to be selected by information gain procedure

Early stopping

	a	b	class
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

- Pre-pruning may stop the growth process prematurely: early stopping
- Classic example: XOR/Parity-problem
  - No individual attribute exhibits any significant association to the class
  - Structure is only visible in fully expanded tree
  - Pre-pruning won't expand the root node
- But: XOR-type problems rare in practice
- And: pre-pruning faster than post-pruning

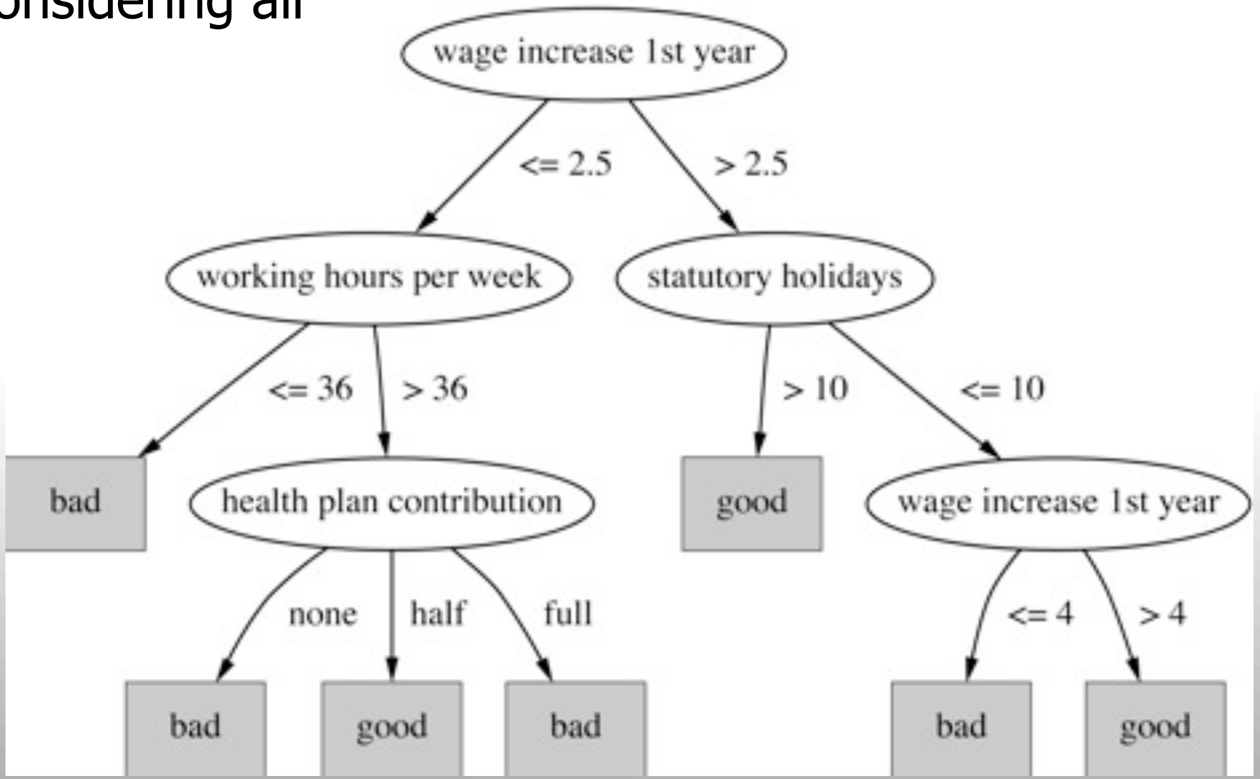
Post-pruning

- First, build full tree
- Then, prune it
  - Fully-grown tree shows all attribute interactions
- Problem: some subtrees might be due to chance effects
- Two pruning operations:

1. **Subtree replacement**
2. Subtree raising

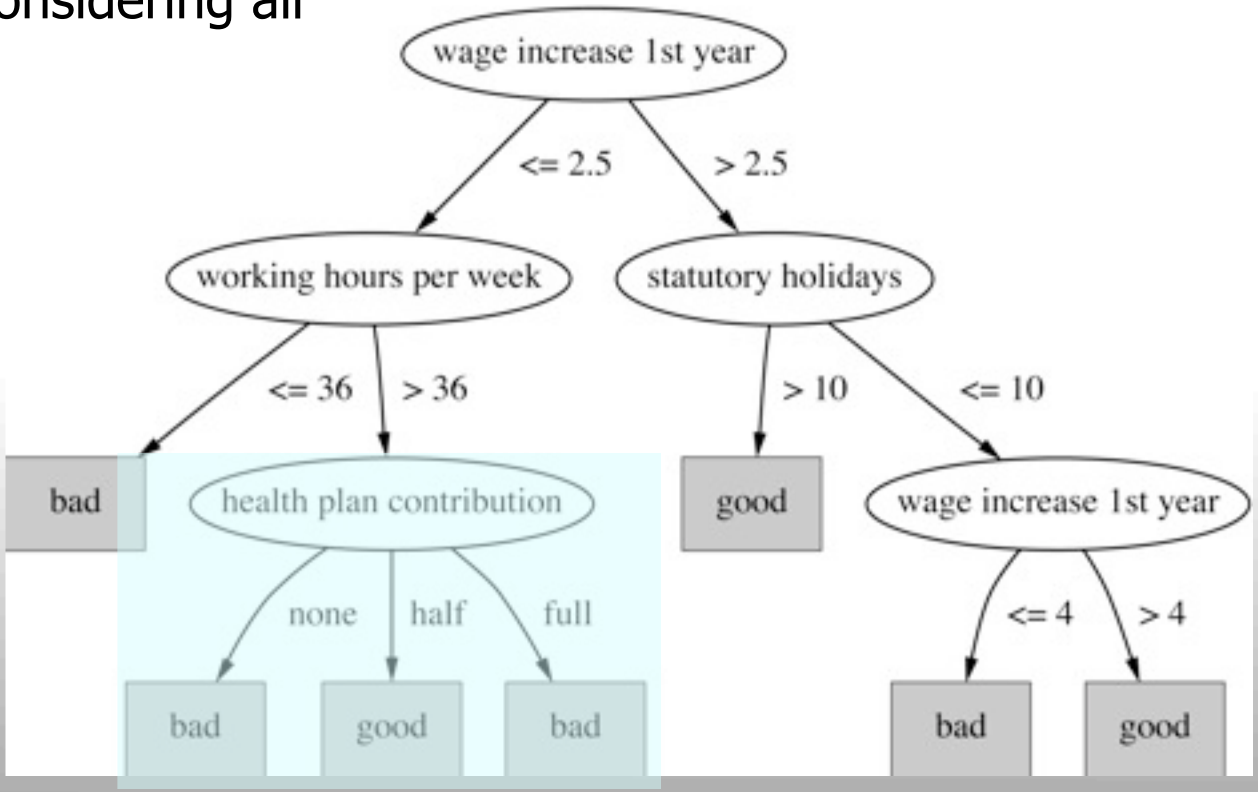
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees



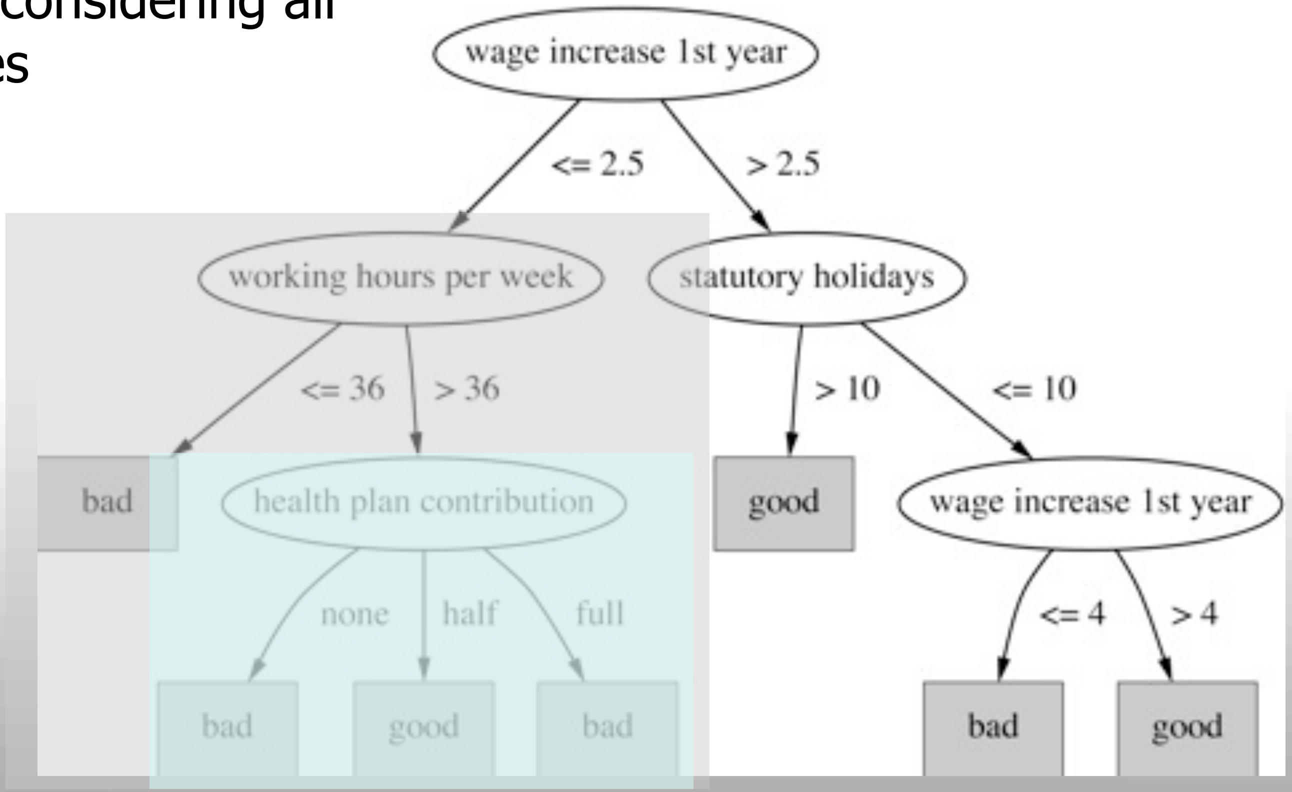
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees



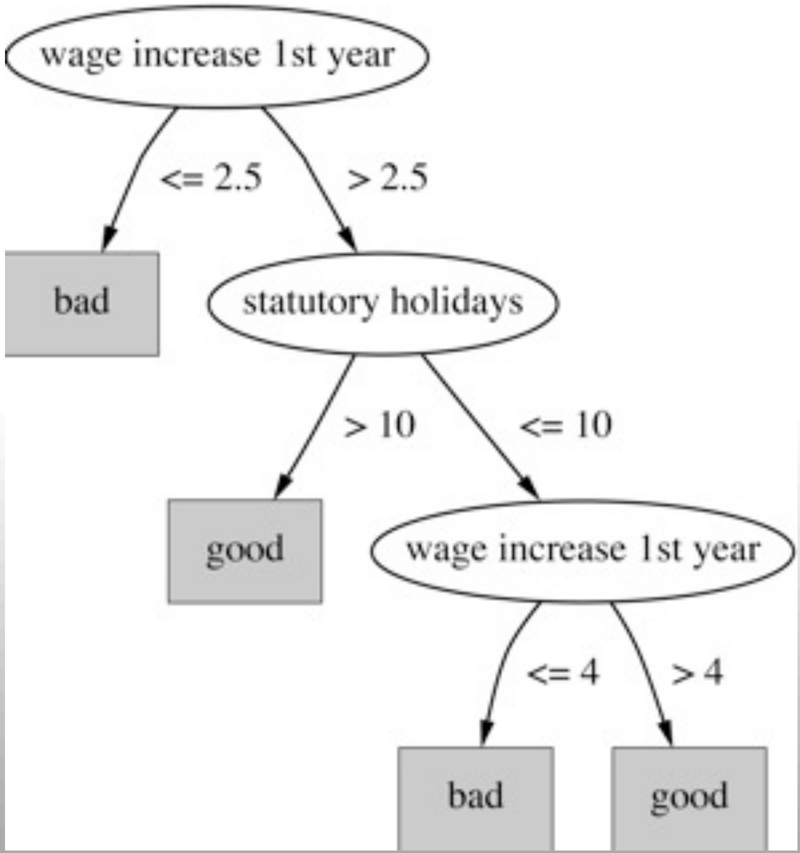
Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees



Subtree replacement

- Bottom-up
- Consider replacing a tree only after considering all its subtrees

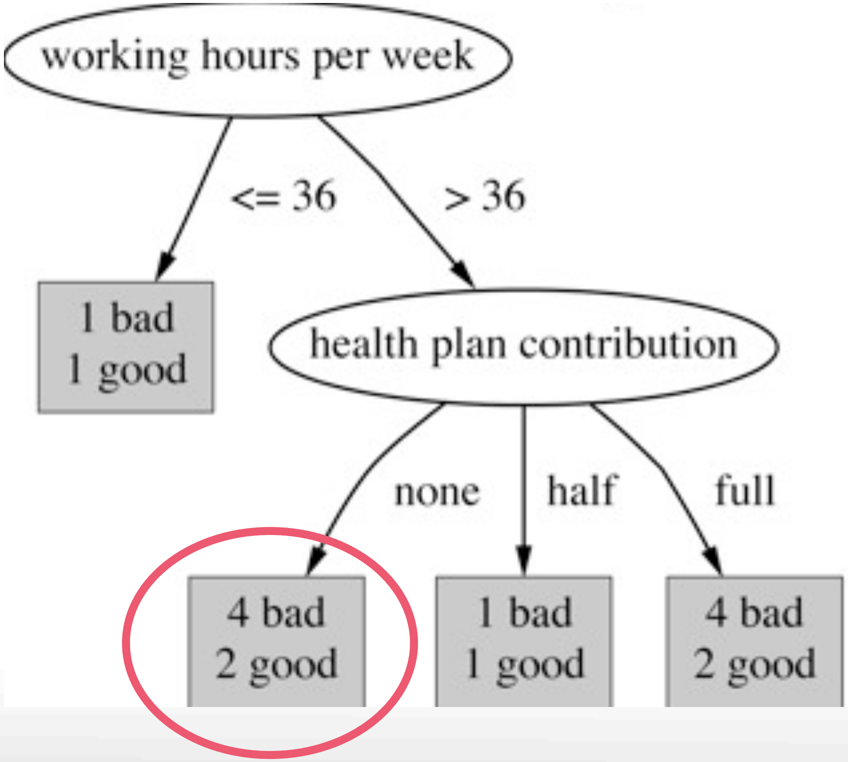


Universiteit Leiden

Estimating error rates

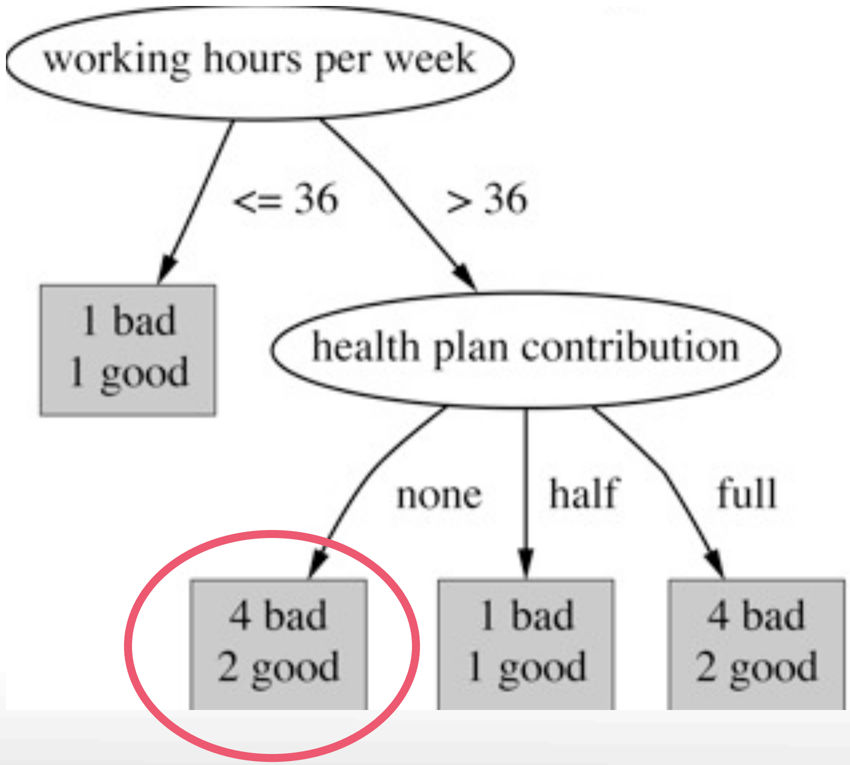
- Prune only if it reduces the estimated error
- Error on the training data is NOT a useful estimator
- Use hold-out set for pruning
  - ("reduced-error pruning")
- C4.5's method
  - Derive confidence interval from **training data**
  - Use a heuristic limit, derived from this, for pruning
  - Standard Bernoulli-process-based method
  - Shaky statistical assumptions (based on training data)

Estimating Error Rates



Estimating Error Rates

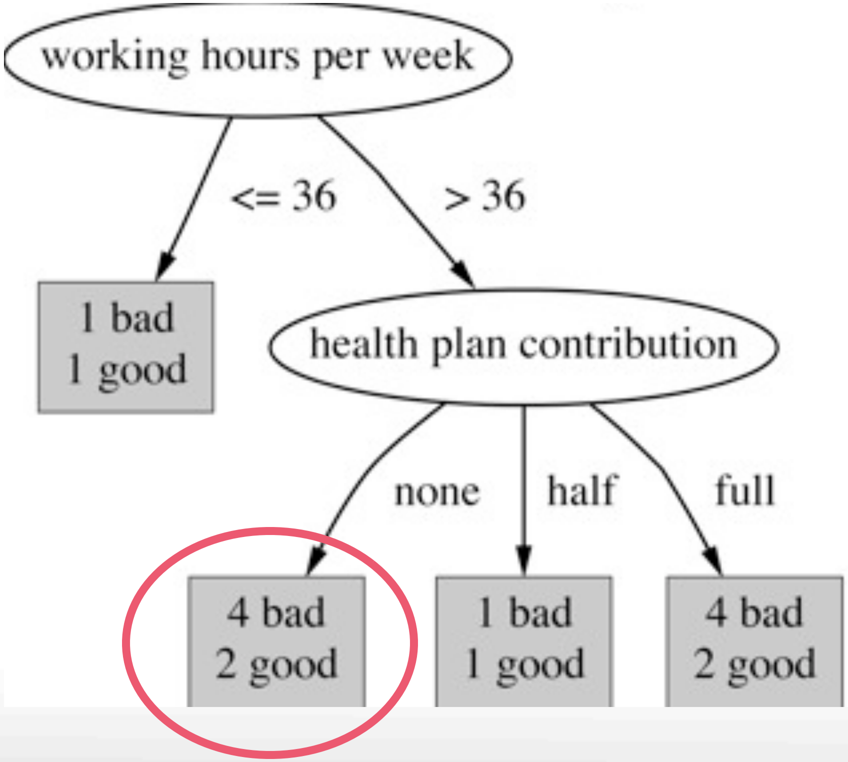
Q: what is the error rate on the training set?



Estimating Error Rates

Q: what is the error rate on the training set?

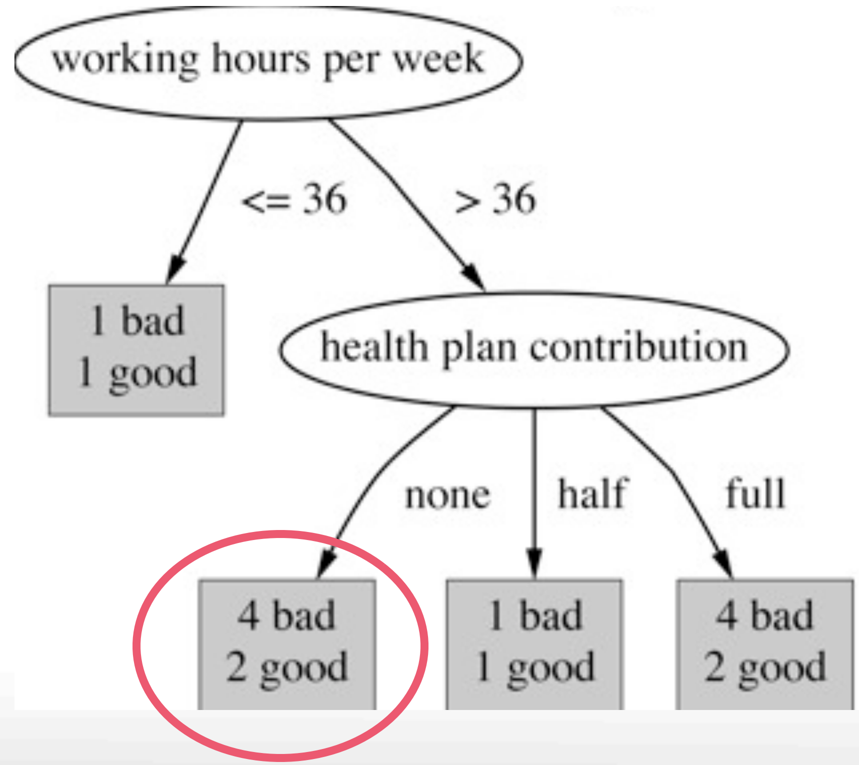
A: 0.33 (2 out of 6)



Estimating Error Rates

Q: what is the error rate on the training set?

A: 0.33 (2 out of 6)

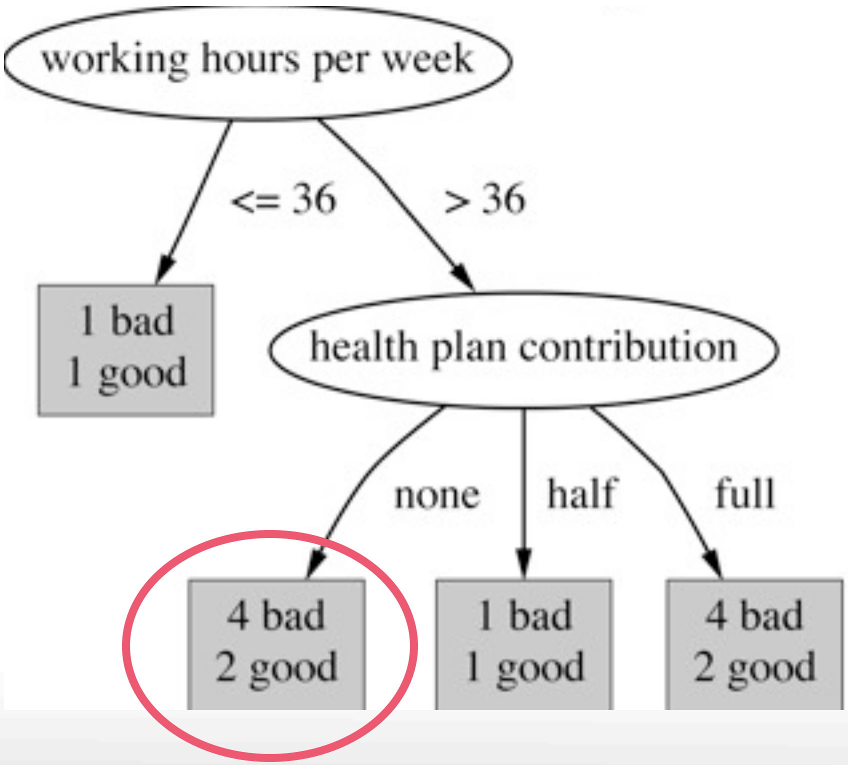


Estimating Error Rates

Q: what is the error rate on the training set?

A: 0.33 (2 out of 6)

Q: will the error on the test set be bigger, smaller or equal?

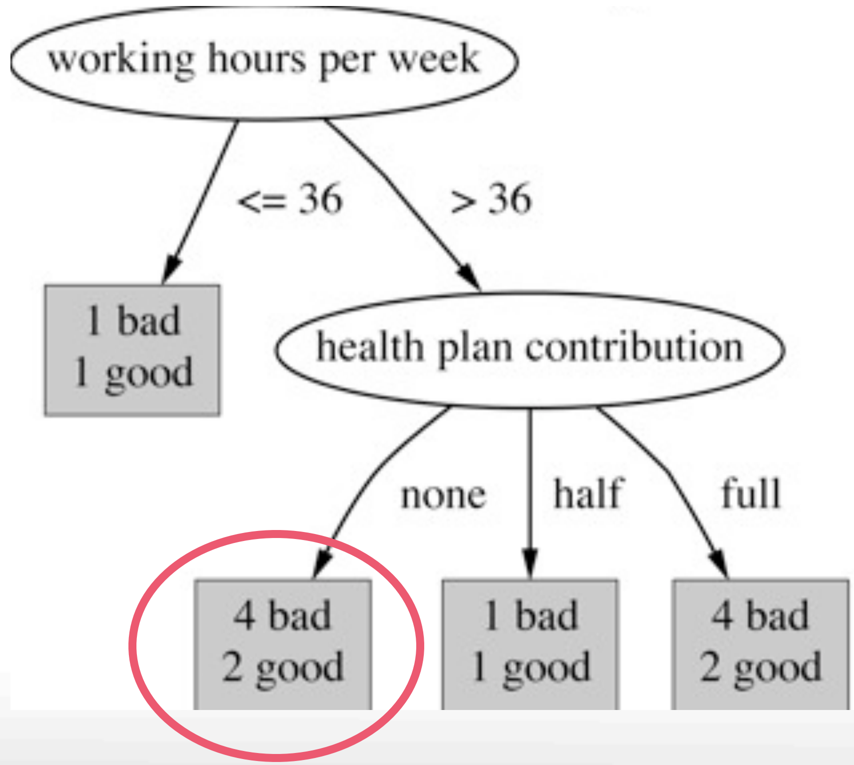


Universiteit Leiden

Estimating Error Rates

Q: what is the error rate on the training set?

A: 0.33 (2 out of 6)

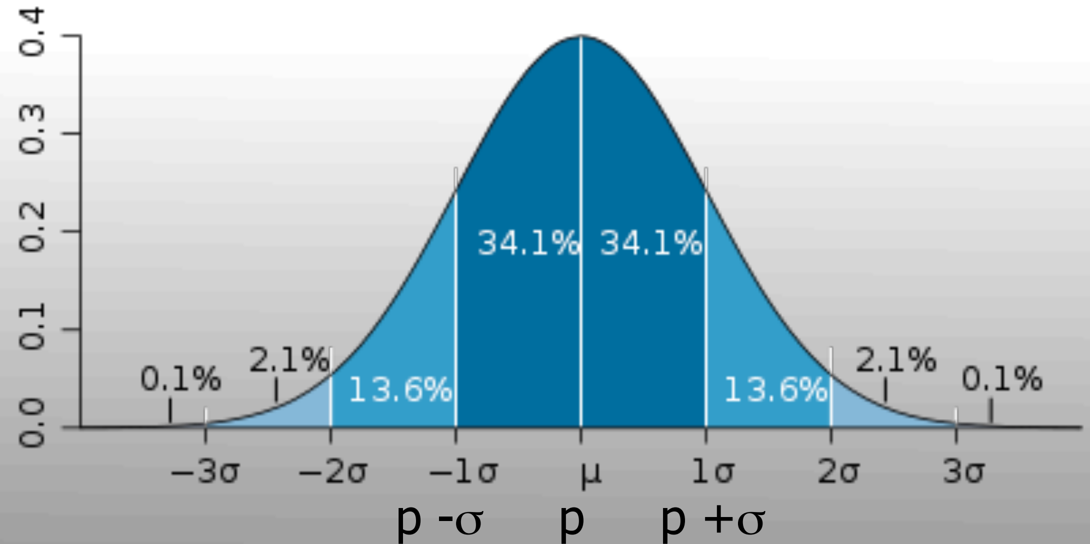


Q: will the error on the test set be bigger, smaller or equal?

A: bigger

Estimating the error

- Assume making an error is Bernoulli trial with probability p
  - p is unknown (true error rate)
- We observe f, the success rate f = S/N
- For large enough N, f follows a Normal distribution
- Mean and variance for f : p, p (1–p)/N



Universiteit Leiden

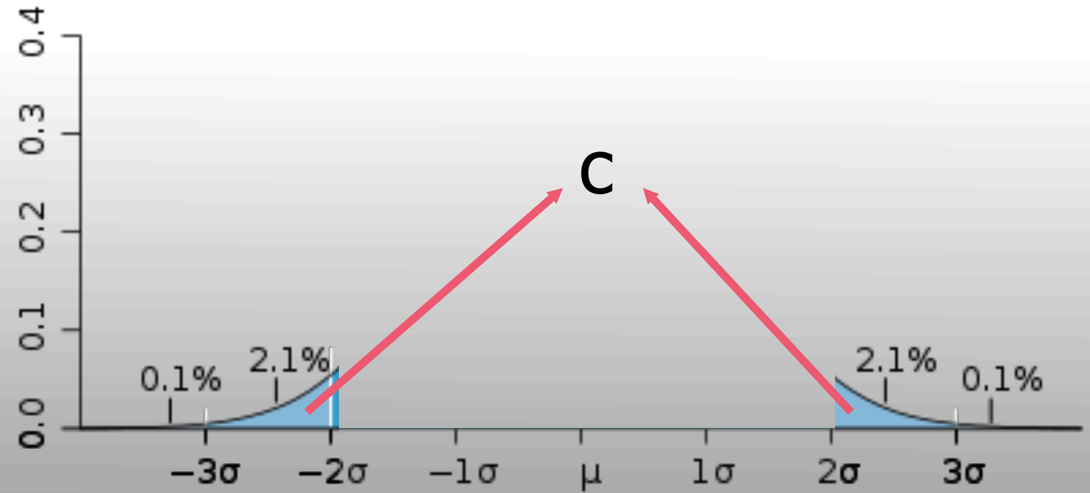
Estimating the error

- c% confidence interval [–z ≤ X ≤ z] for random variable with 0 mean is given by:

$$\Pr[-z \le X \le z] = c$$

- With a symmetric distribution:

$$\Pr[-z \le X \le z] = 1 - 2 \times \Pr[X \ge z]$$



Universiteit Leiden

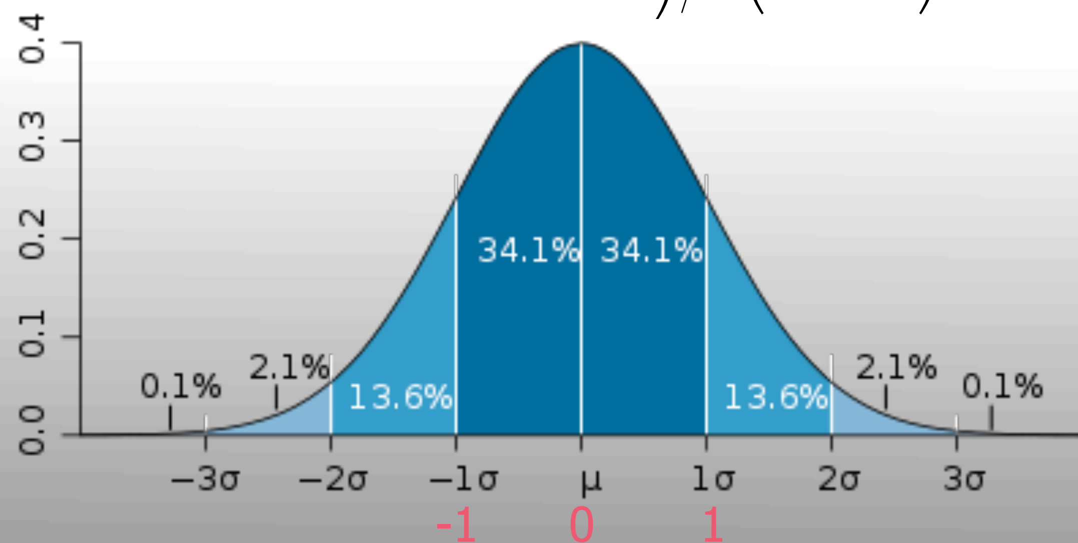
z-transforming f

- Transformed value for f : $\dfrac{f-p}{\sqrt{p(1-p)/N}}$

  (i.e. subtract the mean and divide by the standard deviation)

- Resulting equation: $\Pr\left[-z \le \dfrac{f-p}{\sqrt{p(1-p)/N}} \le z\right] = c$

- Solving for p: $p = \left(f + \dfrac{z^2}{2N} \pm z\sqrt{\dfrac{f}{N} - \dfrac{f^2}{N} + \dfrac{z^2}{4N^2}} \right) \Bigg/ \left(1 + \dfrac{z^2}{N} \right)$



C4.5's method

- Error estimate for subtree is weighted sum of error estimates for all its leaves

- Error estimate for a node (upper bound):

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Bigg/ \left(1 + \frac{z^2}{N} \right)$$

- If c = 25% then z = 0.69 (from normal distribution)

Pr[X ≥ z]	z
1%	2.33
5%	1.65
10%	1.28
20%	0.84
25%	0.69
40%	0.25

C4.5's method

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Bigg/ \left(1 + \frac{z^2}{N} \right)$$

C4.5's method

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Big/ \left(1 + \frac{z^2}{N} \right)$$

f is the observed error

C4.5’s method

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Big/ \left(1 + \frac{z^2}{N} \right)$$

f is the observed error

z = 0.69

C4.5's method

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Big/ \left(1 + \frac{z^2}{N} \right)$$

f is the observed error

z = 0.69

e > f

$e = (f + \varepsilon_1)/(1 + \varepsilon_2)$

C4.5's method

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \bigg/ \left(1 + \frac{z^2}{N} \right)$$

f is the observed error

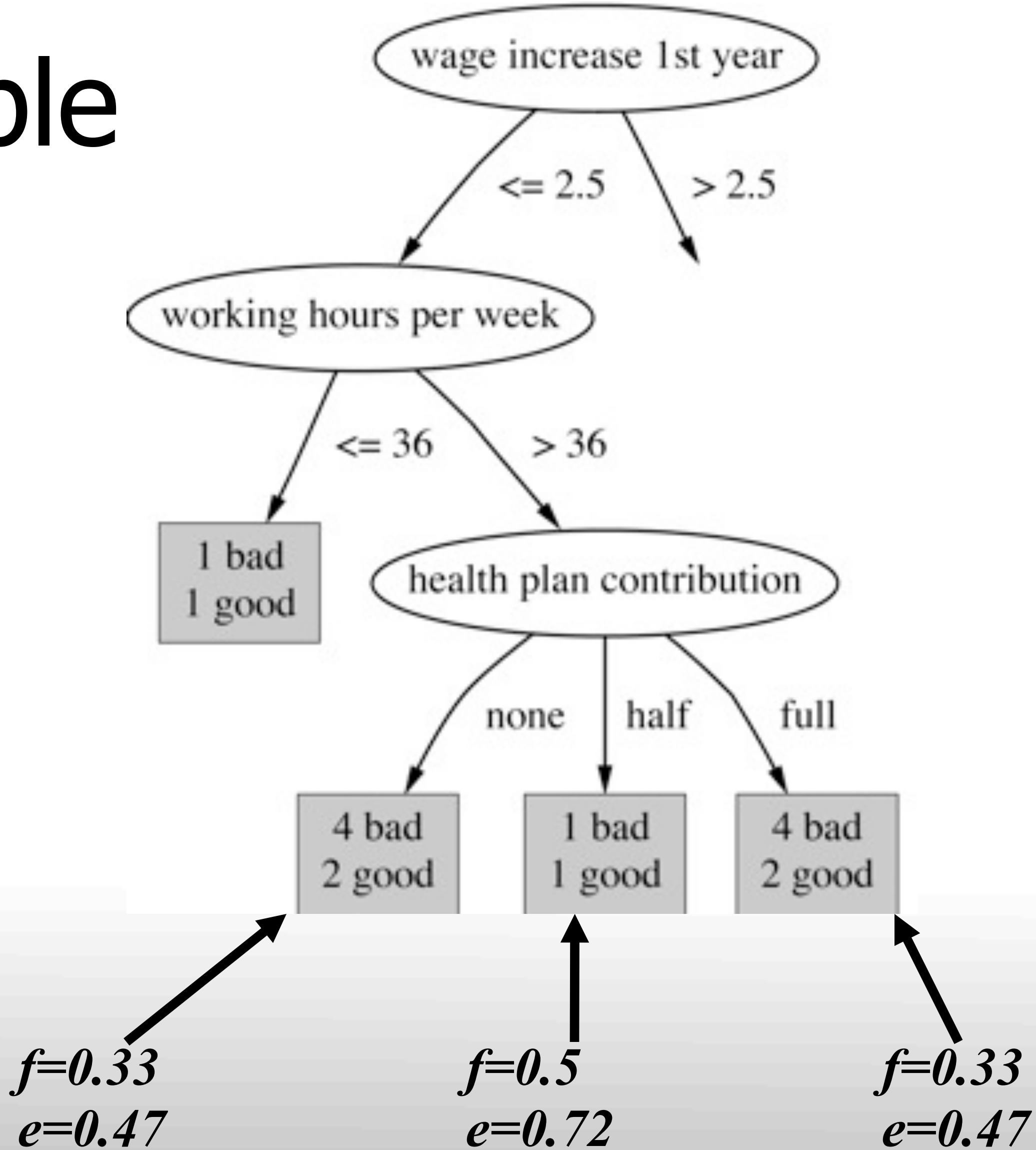
z = 0.69

e > f

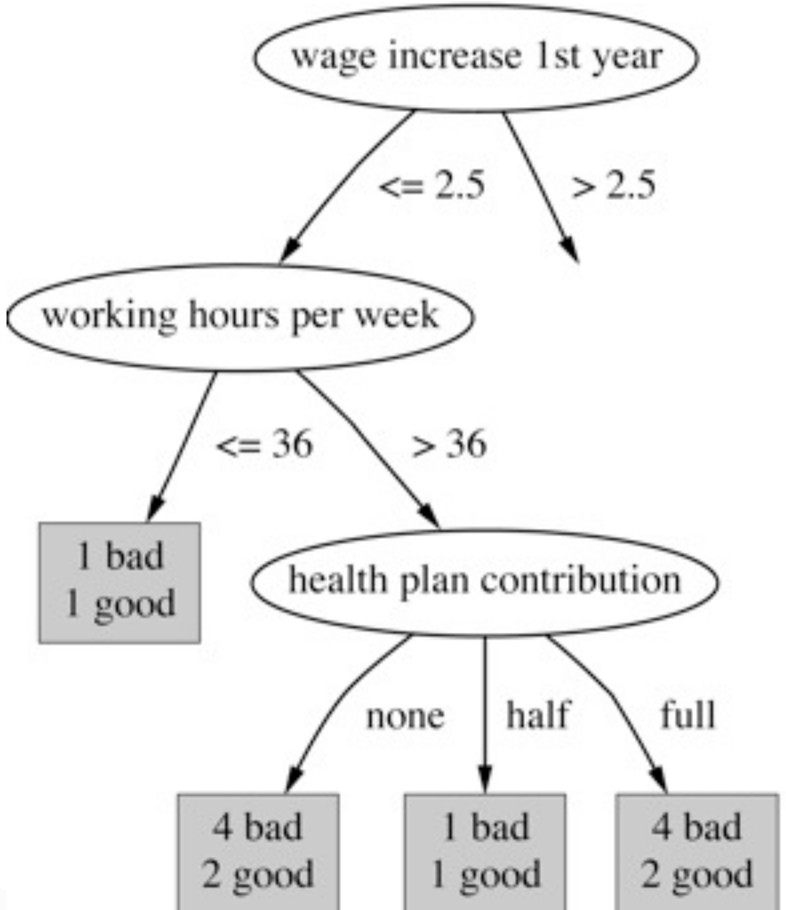
$e = (f + \varepsilon_1)/(1+ \varepsilon_2)$

$N \to \infty$, e = f

Example


wage increase 1st year
<= 2.5
> 2.5
working hours per week
<= 36
> 36
1 bad
1 good
health plan contribution
none
half
full
4 bad
2 good
1 bad
1 good
4 bad
2 good
f=0.33
e=0.47
f=0.5
e=0.72
f=0.33
e=0.47

Example



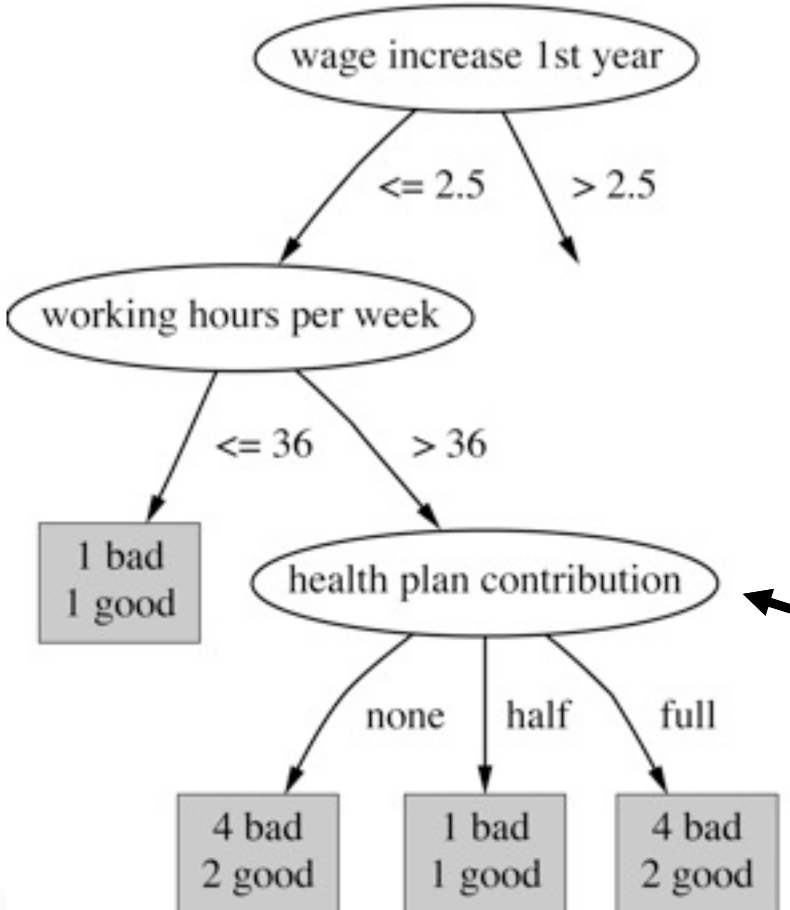
f=0.33
e=0.47

f=0.5
e=0.72

f=0.33
e=0.47

Combined using ratios 6:2:6 gives 0.51

Example



***f* = 5/14=0.36**
***e* = 0.46**
***e* < 0.51**
so prune!

***f*=0.33**
***e*=0.47**

***f*=0.5**
***e*=0.72**

***f*=0.33**
***e*=0.47**

Combined using ratios 6:2:6 gives 0.51

Summary

- Decision Trees
  - splits – binary, multi-way
  - split criteria – information gain, gain ratio, …
  - pruning
- No method is always superior – experiment!