Dataset

attributes

class

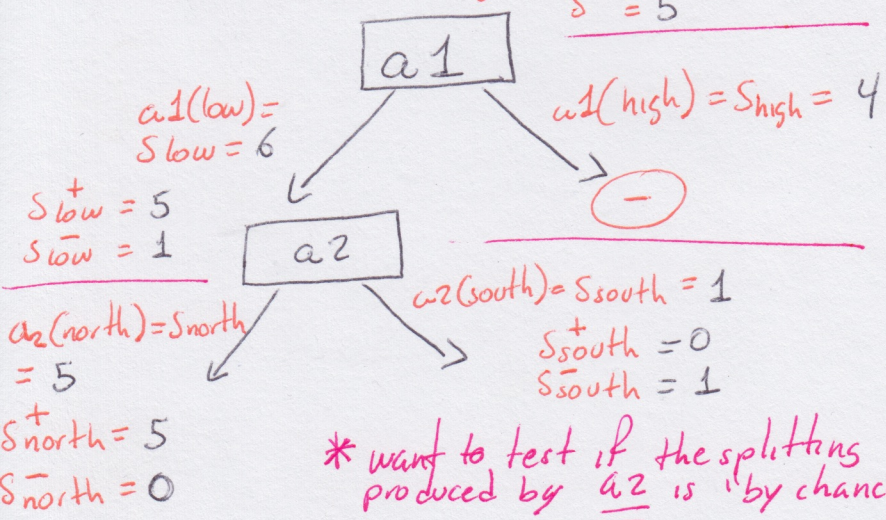
a1	a2	a3	y
high	north	A	−
high	north	B	−
high	south	A	−
high	south	B	−
low	north	A	+
low	north	B	+
low	north	A	+
low	north	B	+
low	north	A	+
low	south	B	−

1st level tree:

$S = \#\text{instances} = 10$

$S^+ = 5$

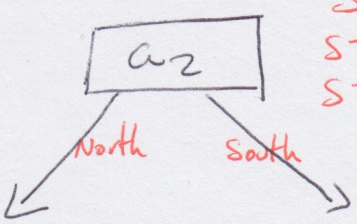
$S^- = 5$



* want to test if the splittings produced by a2 is "by chance"

Zoom into a2 ← simplifying notation

$S = 6$
$S^+ = 5$
$S^- = 1$



$S_{north} = 5$
$S^+_{north} = 5$
$S^-_{north} = 0$

$S_{south} = 1$
$S^+_{south} = 0$
$S^-_{south} = 1$

Hypothesis: a2 is uncorrelated with the decision, then:

S^+_{north} should be <u>close</u> to $\underbrace{S_{north} * \overbrace{S^+/S}^{\text{proportion of } S^+ \text{ in parent}}}_{\text{Expected counts of } S^+_{north} \text{ in the child}}$

S^+_{south} should be close to $S_{south} * S^+/S$

S^-_{north} close to $S_{north} * S^-/S$
S^-_{south} close to $S_{south} * S^-/S$

For the split produced by a2 we need to find the <u>real</u> counts and the <u>expected</u> counts for S^+_{north}, S^+_{south}, S^-_{north}, S^-_{south}

Given $\boxed{Y = +}$ (and a1 = low for the 1st split)

Variable	Real count	Expected
A_2 = north	5	$(5)(5/6) = 25/6$
A_2 = south	0	$(1)(5/6) = 5/6$

$5^+/5$

Given $\boxed{Y = -}$

Snorth

A_2 = north	0	$(5)(1/6) = 5/6$
A_2 = south	1	$(1)(1/6) = 1/6$

Ssouth $5^-/5$

$$C = \sum_{i=1}^{v} \frac{(P_i - P'_i)^2}{P'_i} + \frac{(n_i - n'_i)^2}{n'_i}$$

$i=1$ to v represents all the possible values for the split attribute in this case {north, south}

In this example P_i is the <u>real</u> count of positives (+) with a particular i value {north, or south}

$$C = \frac{(5 - 25/6)^2}{25/6} + \frac{(0 - 5/6)^2}{5/6} \leftarrow i = \text{north}$$

$$+ \frac{(0 - 5/6)^2}{5/6} + \frac{(1 - 1/6)^2}{1/6} = 6$$

$i = \text{south} \rightarrow$

Use chi-square table to figure out if your C is larger than the threshold

Define: confidence level (e.g. 95% = $\alpha = 0.05$

Determine DOF = (num Y labels − 1) *
(num att labels − 1)

In the example DOF = 1 and $\alpha = 0.05$

*Find value in table = 3.8415

if C <= 3.8415 split is by chance and we can prune the node

if C > 3.8415 split is not by chance and we can expand the node

6 > 3.8415 ← the split is relevant