Exercise: calculate Info Gain

- Let’s start with “age”, , see if the entropy gets smaller after using age to split the data.
- Step 1: calculate the entropy of the entire training data set S, which contains 9 positive examples and 5 negative examples.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise: calculate Info Gain

- Let’s start with “age”, , see if the entropy gets smaller after using age to split the data.
- Step 1: calculate the entropy of the entire training data set S, which contains 9 positive examples and 5 negative examples.

$$Entropy(S) = I(9{,}5)$$

$$= \quad \frac{9}{14}\log_2(\frac{9}{14}) \quad \frac{5}{14}\log_2(\frac{5}{14})$$

$$=0.940$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise: calculate Info Gain

- Step 2: count the numbers of positive examples (column p_i) and negative examples (column n_i) in each subset, and then calculate the entropy for each subset, $I(p_i, n_i)$.
- For example, for the "<=30" subset S_1,

$$Entropy(S_1) = I(2,3)$$

$$= \quad \frac{2}{5}\log_2(\frac{2}{5}) \quad \frac{3}{5}\log_2(\frac{3}{5})$$

$$=0.971$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

Similarly,
Entropy(S_2) =0;
Entropy(S_3) = Entropy(S_1)=0.971

Exercise: calculate Info Gain

- Step 3: calculate the weighted average entropy after using age to split the data into three subsets “<=30”, “31..40”, and “>40”.

$$Entropy(age,S)=\frac{5}{14}I(2,3)+\frac{4}{14}I(4,0)+\frac{5}{14}I(3,2)$$

$$=\frac{5}{14}\cdot\ 0.971+\frac{4}{14}\cdot\ 0+\frac{5}{14}\cdot\ 0.971$$

$$=0.694$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

- Class P: buys_computer = “yes”
- Class N: buys_computer = “no”

Exercise: calculate Info Gain

- Step 4: calculate the information gain of using age to split the data into three subsets "<=30", "31..40", and ">40".

$$Gain(age) = Entropy(S) \quad Entropy\ (age, S)$$

$$= 0.940 \quad 0.694 = 0.246$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

Which attribute should be the first node?

- Step 5: repeat the process for each attribute, and then pick the attribute with highest IG as the first node.

$$Gain(age) = 0.246$$

$$Gain(income) = 0.029$$

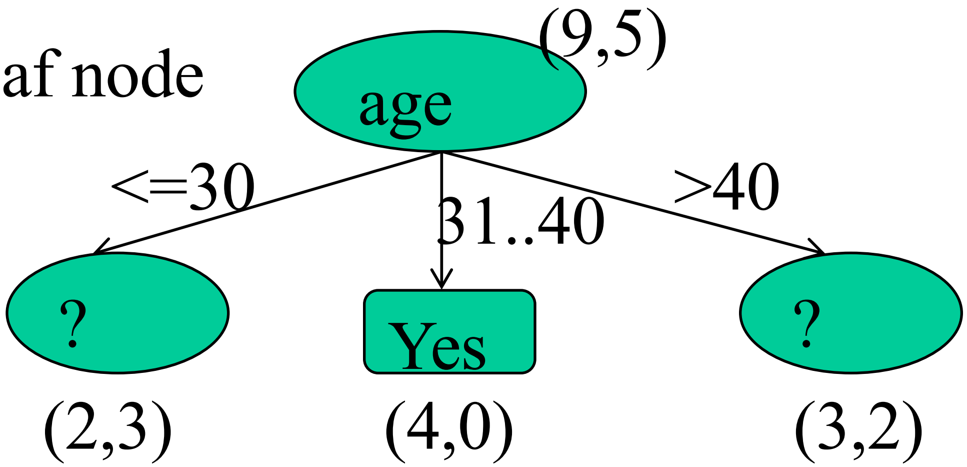
$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

- The DT now hass one leaf node

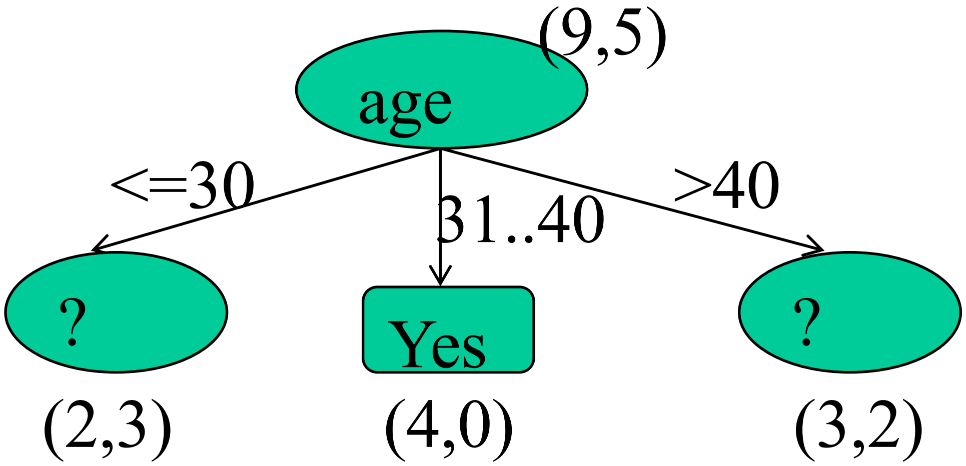
And two subsets that needs

To be further split.



What’s the next step?

- Repeat the prior steps for the subsets (2,3) and (3,2).
  - For subset (2,3), calculate IG for each attribute, pick the attribute with highest IG to replace the question mark.
  - Do the same thing to the subset (3,2)
- Until all nodes are “pure” with all positive examples, or all negative examples.


(9,5)
age
<=30
31..40
>40
?
Yes
?
(2,3)
(4,0)
(3,2)