CISC 372
Advanced Data Analytics
L7 – Decision Tree

https://l1nna.com/372

Last week

- Underfitting vs. Overfitting
- Hyper-parameter tuning & Experimental Protocol
- KDD Process
  - Iterative
  - Data lifecycle
- Data Attributes
  - Numeric/Nominal/Binominal/Ordinal
- Data Types:
  - Relational records
  - Data Metric
  - Document Data
  - Graph Data
  - Structured vs unstructured data
- Data Characteristics
  - Dimensionality/Sparsity
- Data Preprocessing
  - Normalization/Standardization/Encoding/OOV/Discretization/

Last Week

- Ensemble Method
  - Bias-Variance decomposition
  - Bagging
  - Boosting

Evaluating Classification Methods

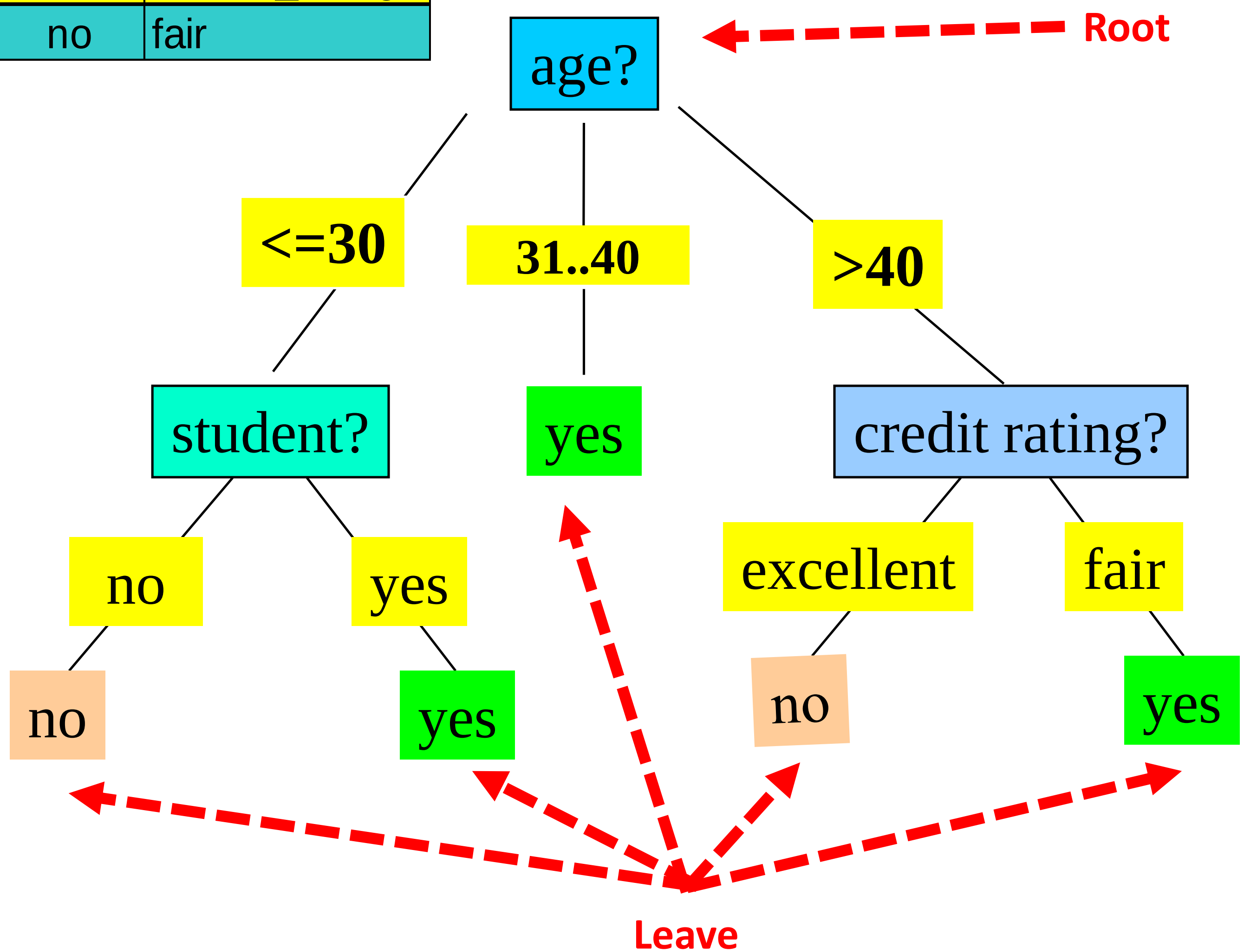
- Accuracy
  - classifier accuracy: predicting class label
  - predictor accuracy: guessing value of predicted attributes
- Speed (efficiency)
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability
  - knowledge and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Decision Tree Induction: Training Dataset

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

Decision Tree for Classification

age	income	student	credit_rating
<=30	high	no	fair

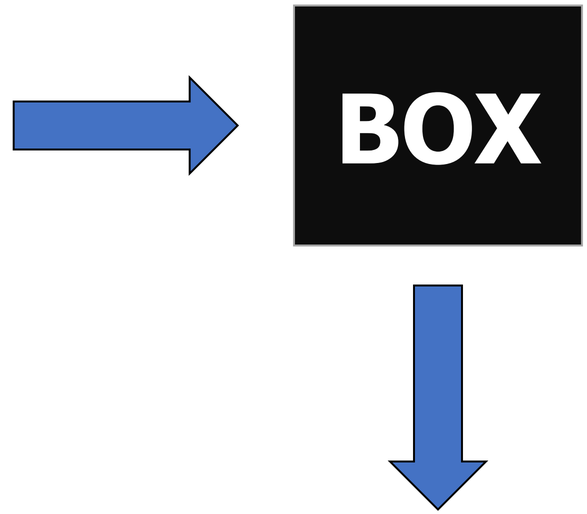


A Magical Black Box

Data table

Class Attribute

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

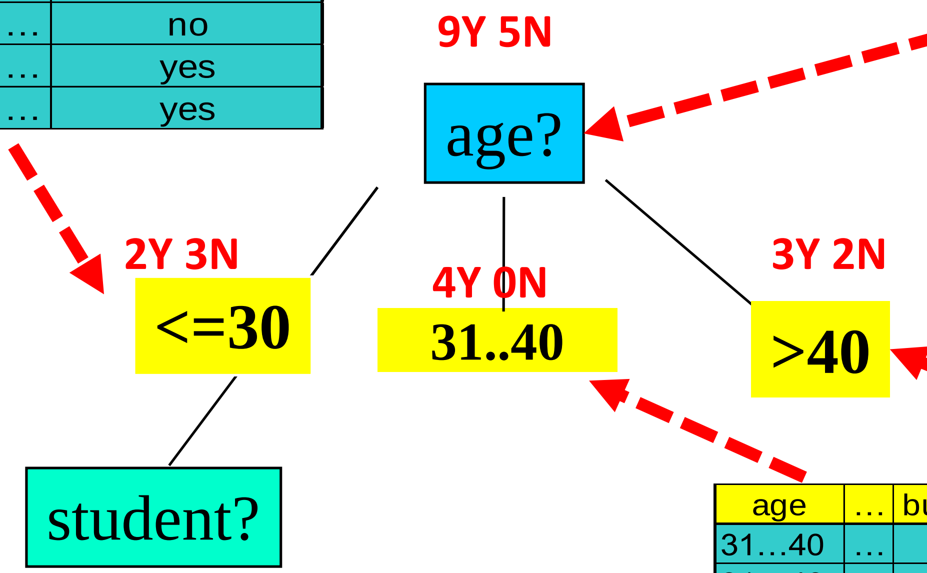


Age is the best attribute to create a node in the tree.

Output: A Decision Tree for “*buys_computer*”

BOX

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes



BOX

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no

age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

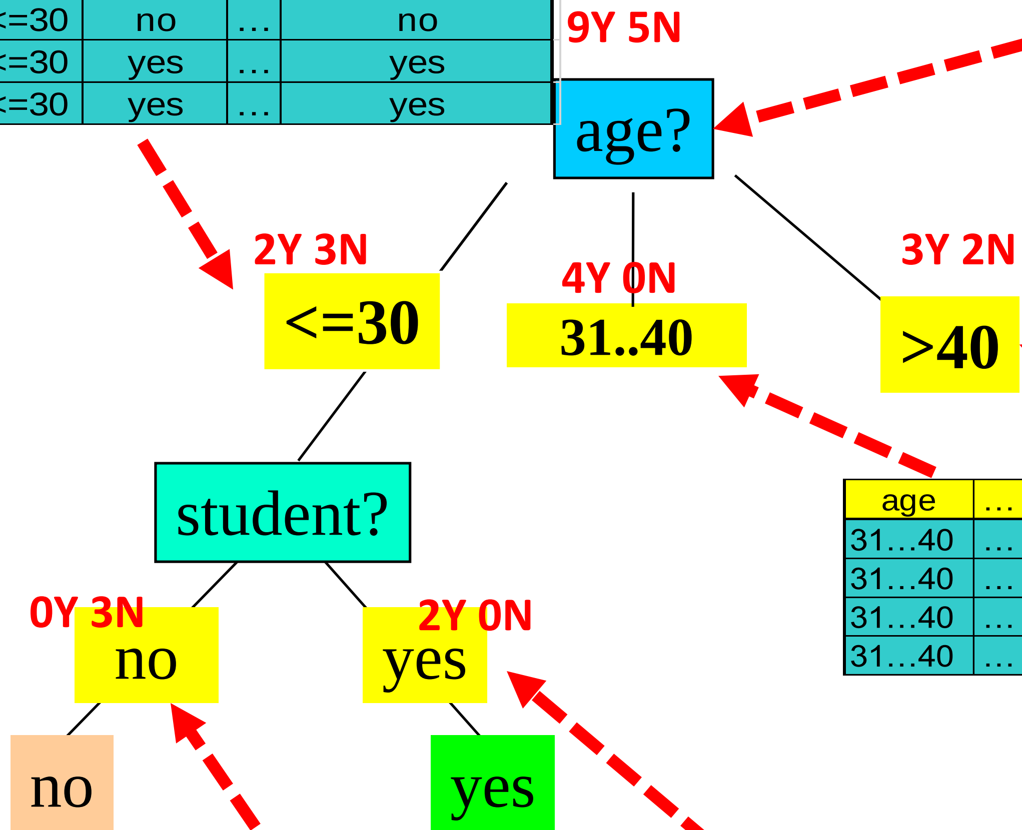
Output: A Decision Tree for "*buys_computer*"

BOX

age	student?	…	buys_computer
<=30	no	…	no
<=30	no	…	no
<=30	no	…	no
<=30	yes	…	yes
<=30	yes	…	yes

BOX

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no



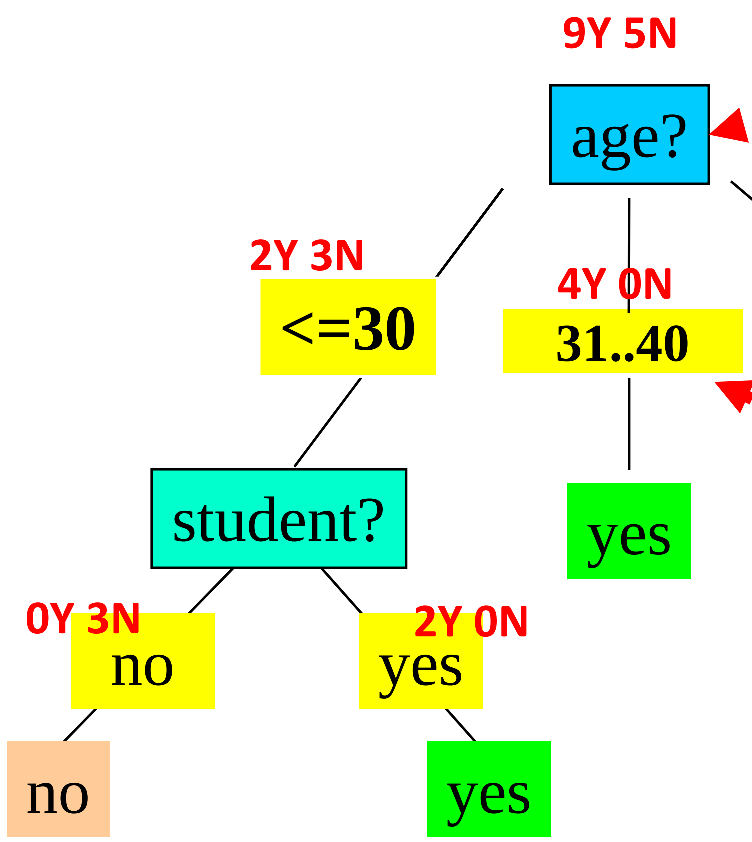
age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

age	student?	buys_computer
<=30	no	no
<=30	no	no
<=30	no	no

age	student?	buys_computer
<=30	yes	yes
<=30	yes	yes

Output: A Decision Tree for “*buys_computer*”



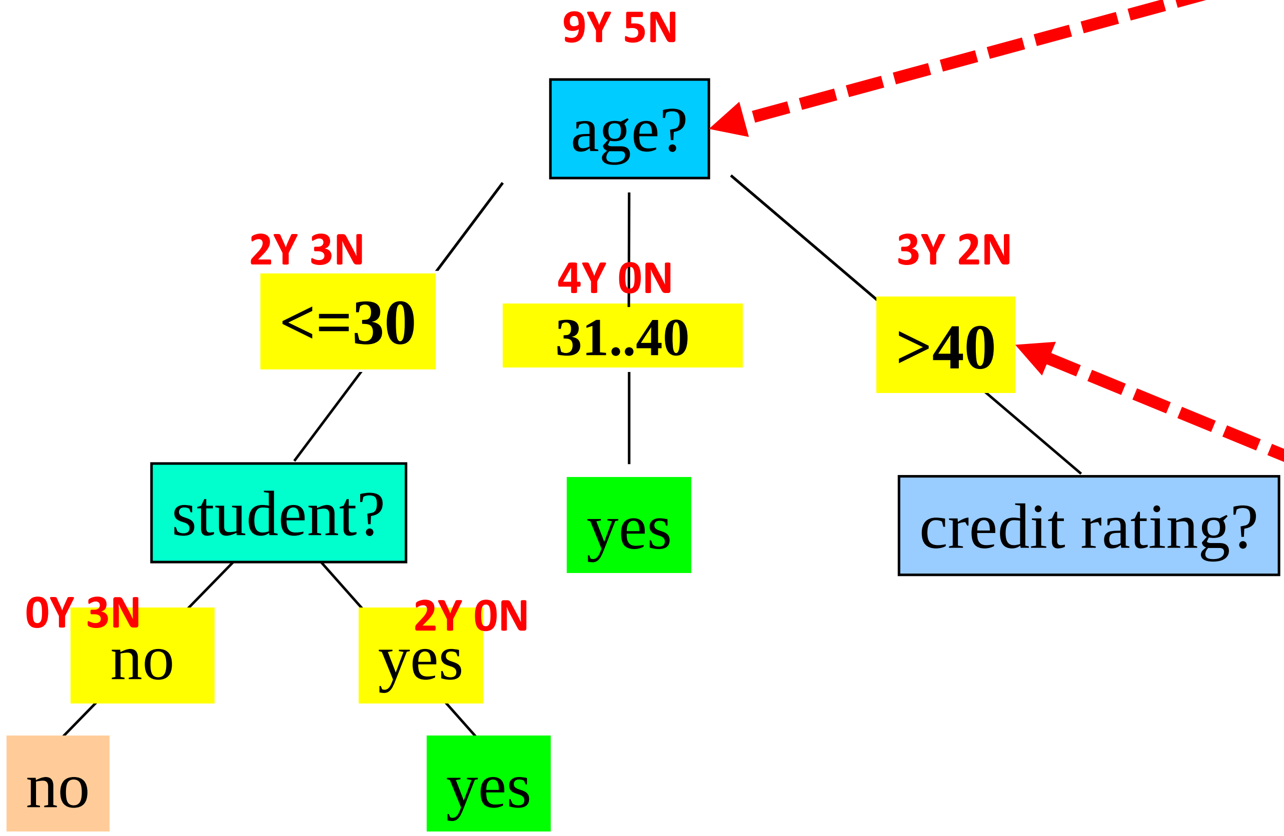
BOX

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no

age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

Output: A Decision Tree for “*buys_computer*”



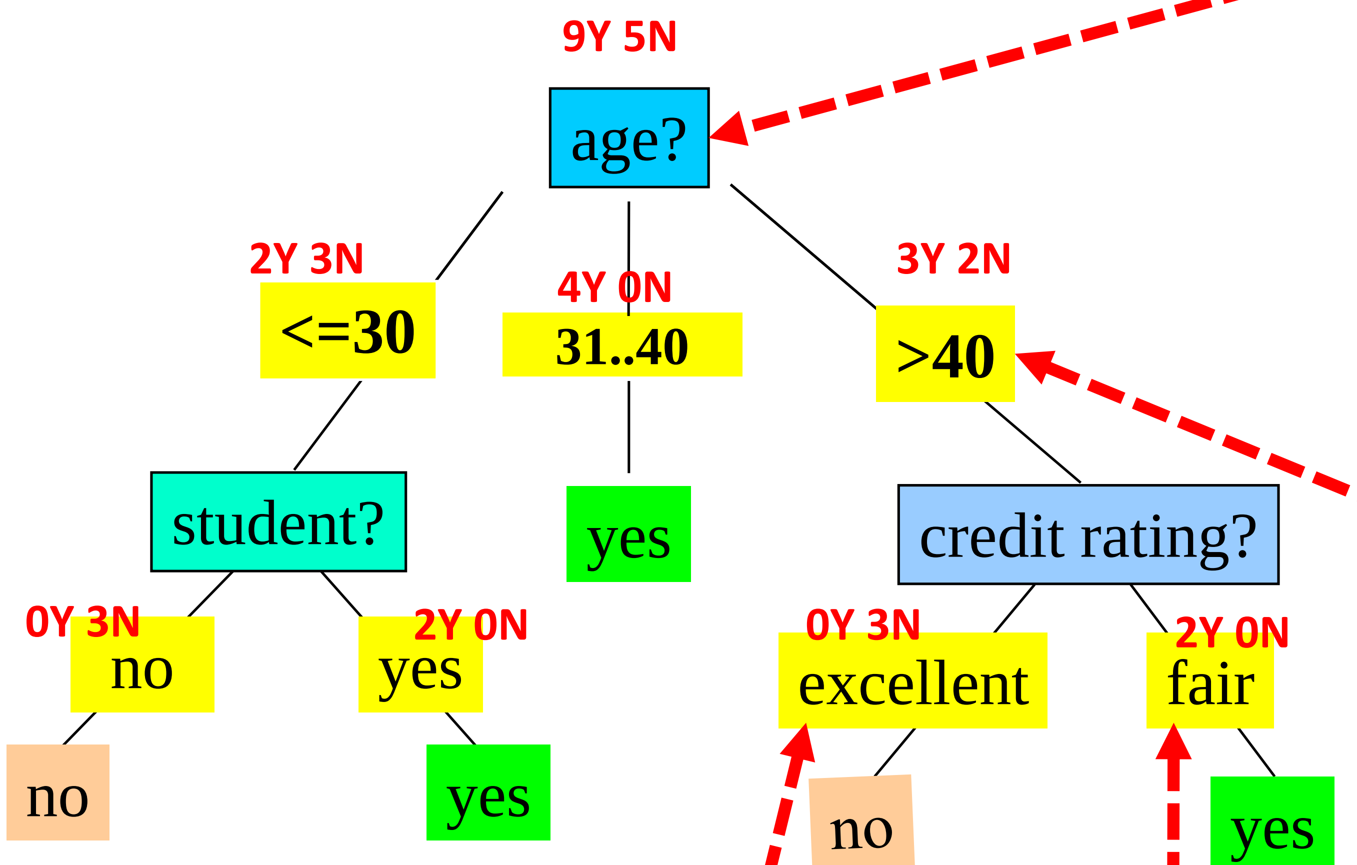
BOX

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no

BOX

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

Output: A Decision Tree for “*buys_computer*”



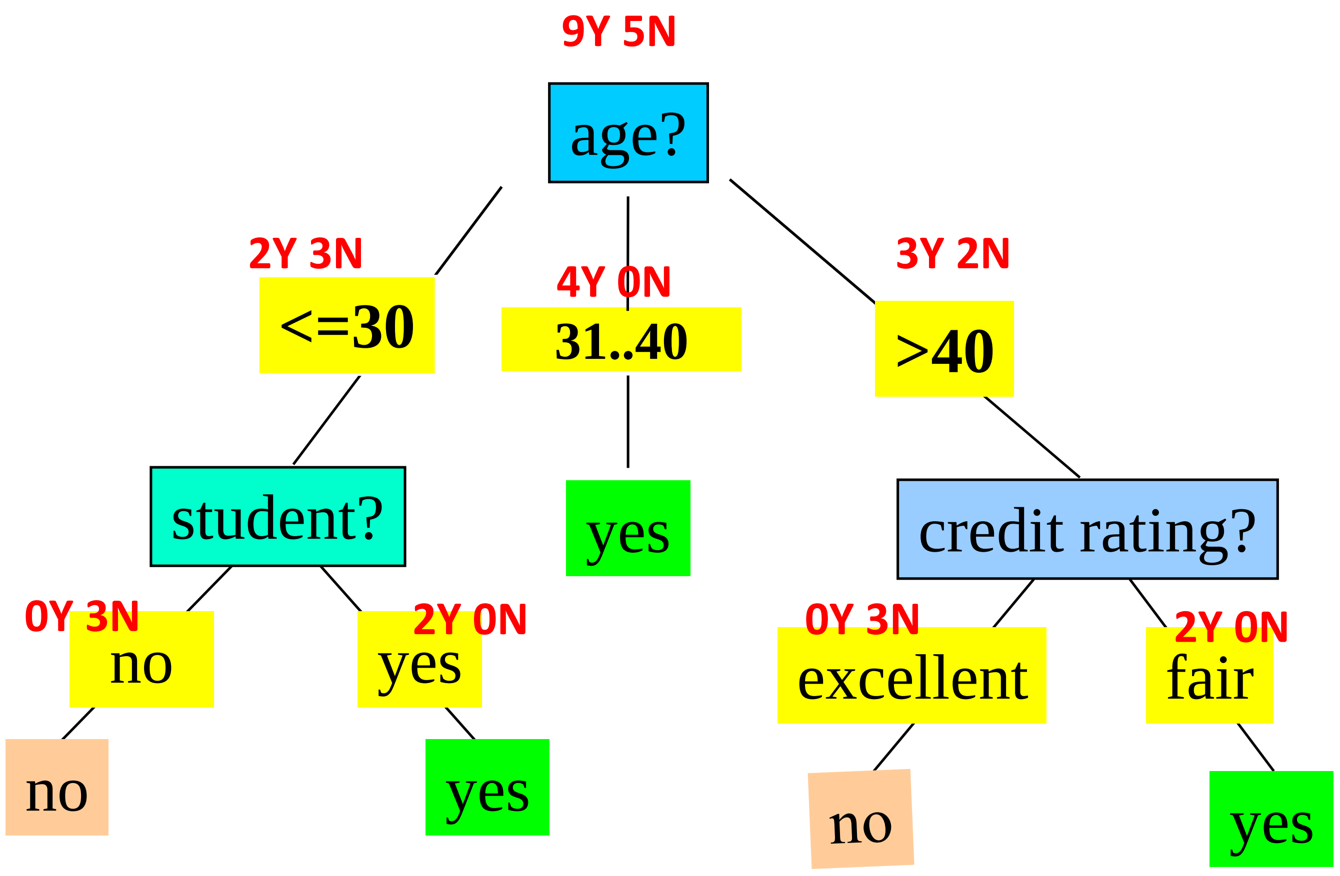
age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no

age	credit	…	buys_computer
>40	excellent	…	no
>40	excellent	…	no
>40	excellent	…	no
>40	fair	…	yes
>40	fair	…	yes

age	credit	buys_computer
>40	excellent	no
>40	excellent	no
>40	excellent	no

age	credit	buys_computer
>40	fair	yes
>40	fair	yes

Output: A Decision Tree for *"buys_computer"*



Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Training examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

BOX

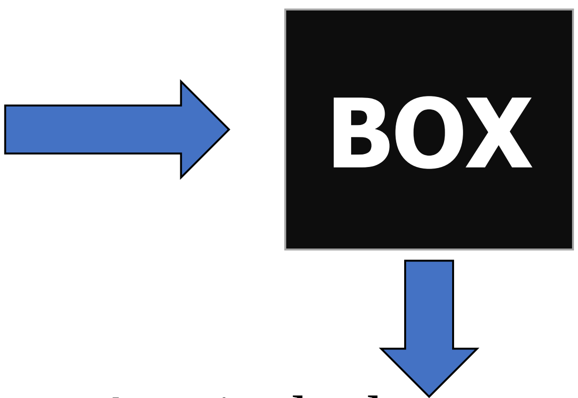
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

What is in the box?

Data table

Class Attribute

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



Age is the best attribute to create a node in the tree.

The chosen attribute should carry **more information than** the others w.r.t. the **class attribute**.

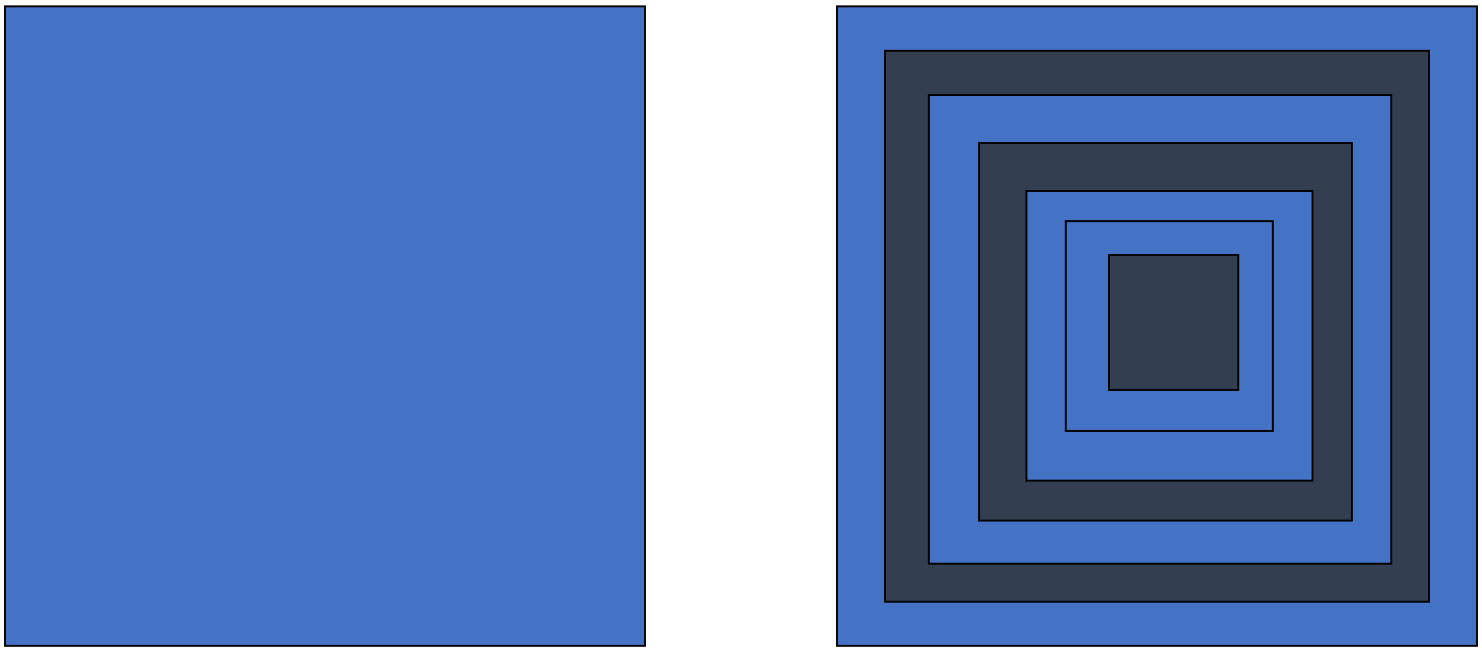
Then how can we measure information?

Entropy – a measure of disorder

Which one has higher entropy?

Source: http://www.organizingfabulously.com/blog/2015/03/02/messy-vs-tidy-desks

Which one has higher entropy?



Which one has higher entropy?

buys_computer
yes
no
yes
yes
no
no
yes
yes
yes
no
no
no
yes
yes

buys_computer
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes

Which one has higher entropy?

buys_computer
yes
no
yes
yes
no
no
yes
yes
yes
no
no
no
yes
yes

buys_computer
no
no
no
yes
yes
yes
yes
yes
yes
yes
yes
no
yes
no

We need a quantitative measure of information (i.e. disorder).

Logarithm

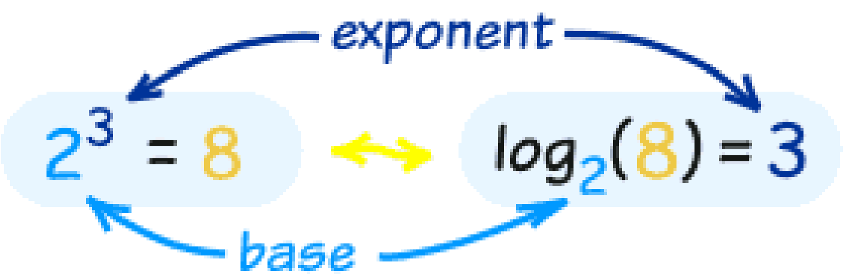
How many of *one number* do we multiply to get *another number?*

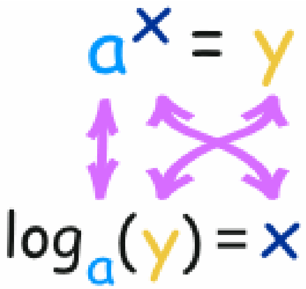
Example: How many 2s do we multiply to get 8?

$$2^{?} = 8$$

Answer: **2 × 2 × 2 = 8**, so we needed to multiply 3 of the **2**s to get **8**

So the logarithm to base 2 of 8 is 3





$$10^{?} = 10000$$

$$\log_{10}(10000) = 4 = \log(10000)$$

$$\log_2(32) = \frac{\log(32)}{\log(2)} = 5$$

$$log_x(y) = \frac{log_{10}(y)}{log_{10}(x)} = \frac{log\ (y)}{log(x)}$$

$$log(1) = 0$$

$log(0)$ non-existed

Entropy – a measure of disorder (impurity in class attribute)

Information needed (i.e. entropy) to classify a random record in D

m is all the unique values for the *class attribute*

The probability of seeing *i*

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Loop for each unique value *i*

Binominal class attribute: '*Yes*' or '*No*'

Let:

- *x denotes the count of 'Yes'*
- *y denotes the count of 'No'*

$$p('yes') = \frac{x}{x+y} \qquad p('no') = \frac{y}{x+y}$$

$$Info(D) = -\left[p('yes')\ log_2\left(p('yes')\right) + p('no')\ log_2\left(p('no')\right)\right]$$

$$Info(D) = I(x,y) = -\frac{x}{x+y}\ log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\ log_2\left(\frac{y}{x+y}\right)$$

Entropy – a measure of disorder (impurity in class attribute)

buys_computer
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes
yes

buys_computer
no
no
no
no
no
no
no
no
no
no
no
no
no
no

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info(D) = I(x,y)$$
$$= -\frac{x}{x+y} log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y} log_2\left(\frac{y}{x+y}\right)$$

$$x = 14, y = 0$$

$$Info(D) = I(14{,}0) = -\frac{14}{14}\log_2(\frac{14}{14}) - \frac{0}{14}\log_2(\frac{0}{14}) = -1\times 0 - 0\times 0 = 0$$

Entropy – a measure of disorder (impurity in class attribute)

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info(D) = I(x,y) = -\frac{x}{x+y}\,log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\,log_2\left(\frac{y}{x+y}\right)$$

$$Info(D) = I(7,7) = -\frac{7}{14}\log_2(\frac{7}{14}) - \frac{7}{14}\log_2(\frac{7}{14})$$

$$=(-0.5\times\log_2 0.5) - (0.5\times\log_2 0.5) = (-0.5\times -1) - (0.5\times -1) = 1$$

buys_computer
yes
no
yes
yes
no
no
yes
yes
yes
no
no
no
yes
no

buys_computer
no
no
no
yes
yes
yes
yes
yes
yes
yes
yes
no
yes
no

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14})$$

$$= (-0.643\times\log_2 0.643) - (0.357\times\log_2 0.357)$$

$$= (-0.643\times -0.637) - (0.357\times -1.485)$$

$$=0.410 - (-0.531) = 0.940$$

Conditional Entropy (Weighted Average)

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes

age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

v is all the unique values for the conditional attribute/feature *A*

Size of the table of value *j*

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Information for the table of value *j*

Loop through each unique value for the conditional attribute value *j*

Total size of all tables

Conditional Entropy (Weighted Average)

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes

age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

age	yes	no	total	I(yes,no)
<=30	2	3	5	I(2,3)=0.971
31…40	4	0	4	I(4,0)=0
>40	3	2	5	I(2,3)=0.971

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Information Gain

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes

age	…	buys_computer
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes

age	…	buys_computer
>40	…	no
>40	…	no
>40	…	no
>40	…	yes
>40	…	yes

Entropy (information need) before splitting

Entropy (information needed) after splitting with attribute *A*

$$Gain(A) = \underbrace{Info(D) - Info_A(D)}$$

The amount of reduced entropy if we split on attribute *A*.

i.e. the amount of information we can gain from attribute age w.r.t. buys_computers.

age	…	buys_computer
<=30	…	no
<=30	…	no
<=30	…	no
<=30	…	yes
<=30	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
31…40	…	yes
>40	…	yes
>40	…	yes
>40	…	no
>40	…	yes
>40	…	no

Information Gain

age	yes	no	total	I(yes,no)
Any	9	5	14	0.94
<=30	2	3	5	0.971
31…40	4	0	4	0
>40	3	2		0.971

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14})$$

$$= 0.410 - (-0.531) = 0.940$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info(D) = I(x,y)$$

$$= -\frac{x}{x+y}\, log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y} log_2\left(\frac{y}{x+y}\right)$$

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

$$+ \frac{5}{14} I(3,2) = 0.694$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

What is in the box?

Data table

Class Attribute

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

BOX

- Calculate information gain for each feature.
- Pick the feature that has highest information gain.

A Decision Tree for “*buys_computer*” (1st level)

$$I(x,y) = -\frac{x}{x+y}\log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\log_2\left(\frac{y}{x+y}\right)$$

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means “age <=30” has 5 out of 14 samples, with 2 yes’es and 3 no’s.

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$$Gain(income) = 0.029$$

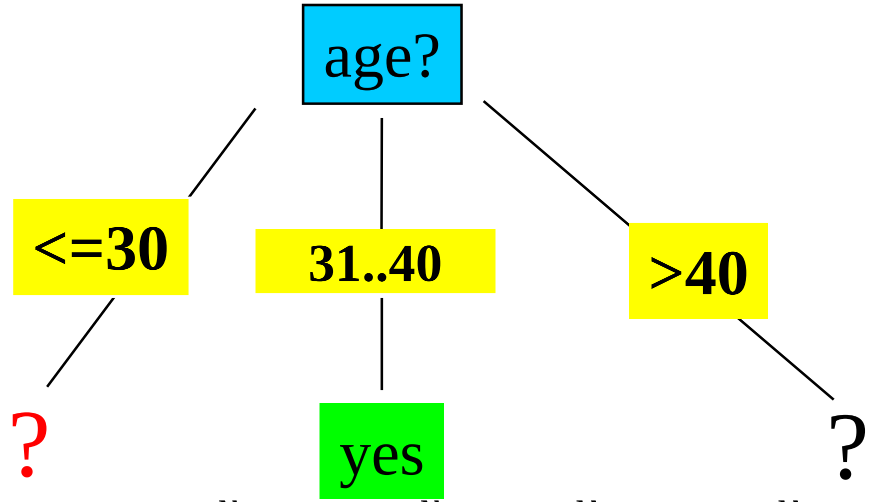
$$Gain(student) = 0.151$$

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

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

age	yes	no	I(yes,no)
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

A Decision Tree for "*buys_computer*" (2nd level)



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

$$I(x,y) = -\frac{x}{x+y}\log_2\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\log_2\left(\frac{y}{x+y}\right)$$

$$Info(D[age <= 30])$$

$$= I(2,3) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

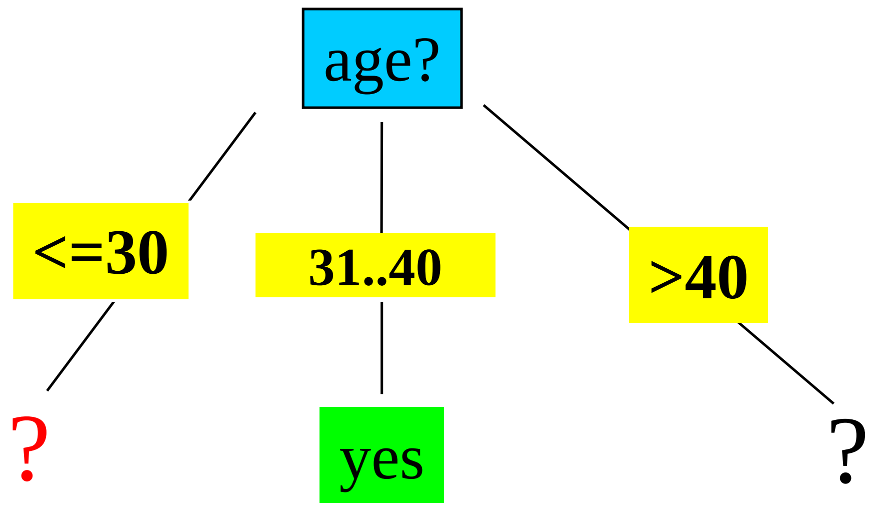
$$Info_{income}(D[age <= 30])$$

$$= \frac{1}{5}I(1,0) + \frac{2}{5}I(1,1) + \frac{2}{5}I(0,2)$$

$$= 0 + 0.4 + 0 = 0.4$$

income	yes	no	I(yes,no)
low	1	0	0
medium	1	1	1
high	0	2	0

A Decision Tree for "*buys_computer*" (2nd level)



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

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

$$Info_{student}(D[age <= 30]) = 0$$

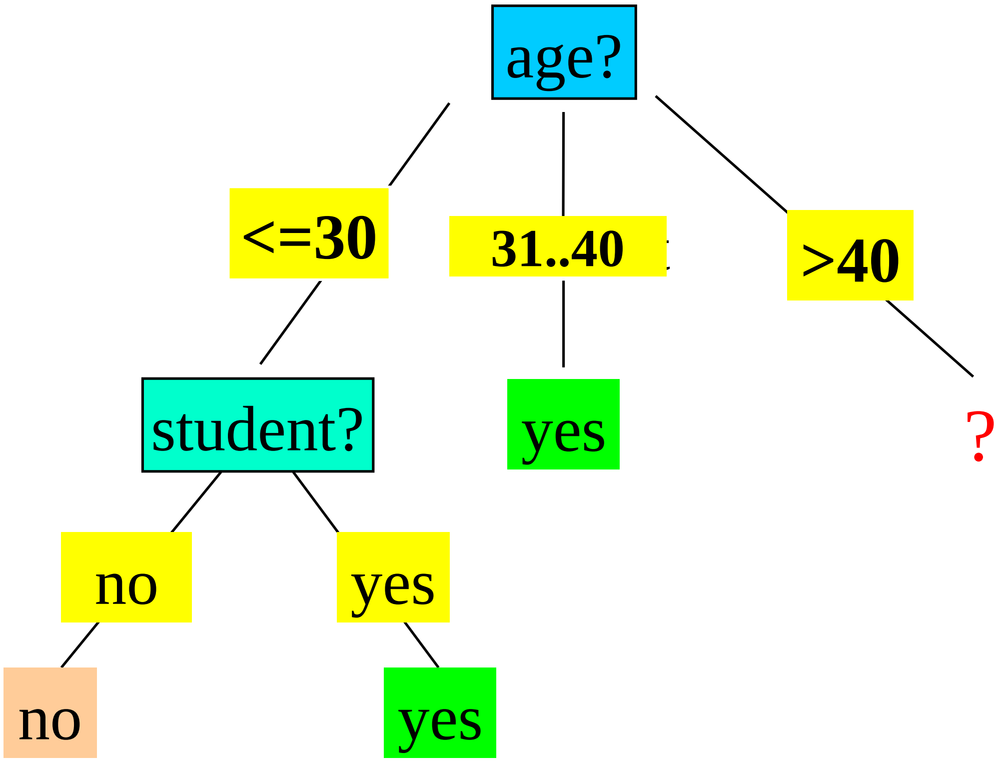
$$Info_{credit\_rating}(D[age <= 30])$$

$$= \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1) = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 1 = 0.951$$

student	yes	no	I(yes,no)
yes	2	0	0
no	0	3	0

credit_rating	yes	no	I(yes,no)
fair	1	2	0.918
exceller	1	1	1

A Decision Tree for “*buys_computer*” (2nd level)



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

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

$$Info_{income}(D[age <= 30]) = 0.4$$

$$Info_{student}(D[age <= 30]) = 0$$

$$Info_{credit\_rating}(D[age <= 30]) = 0.951$$

$$Gain(income) = 0.971 - 0.4 = 0.571$$

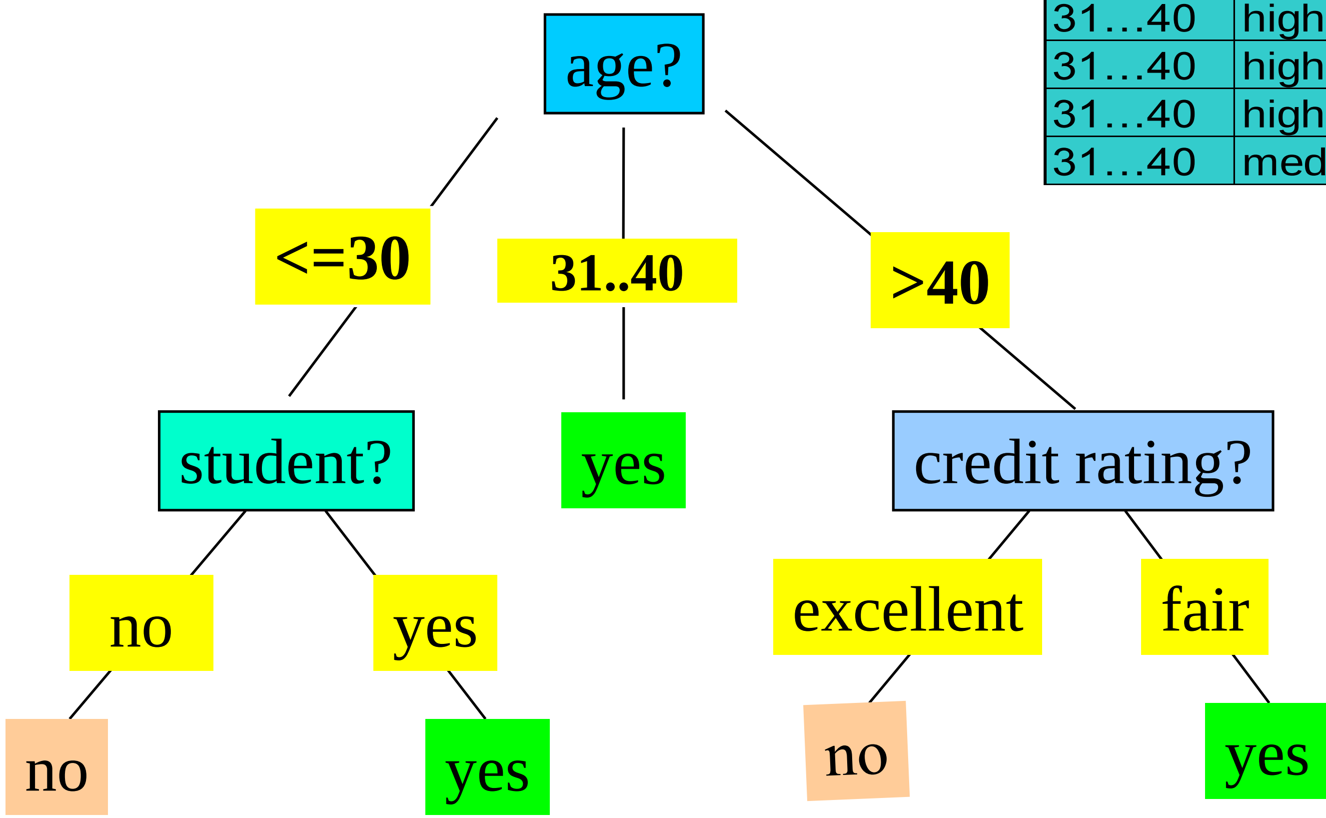
$$Gain(student) = 0.971 - 0 = 0.971$$

$$Gain(credit\_rating) = 0.971 - 0.951 = 0.02$$

Overfitting and Tree Pruning

- Overfitting: An induced tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Poor accuracy for unseen samples

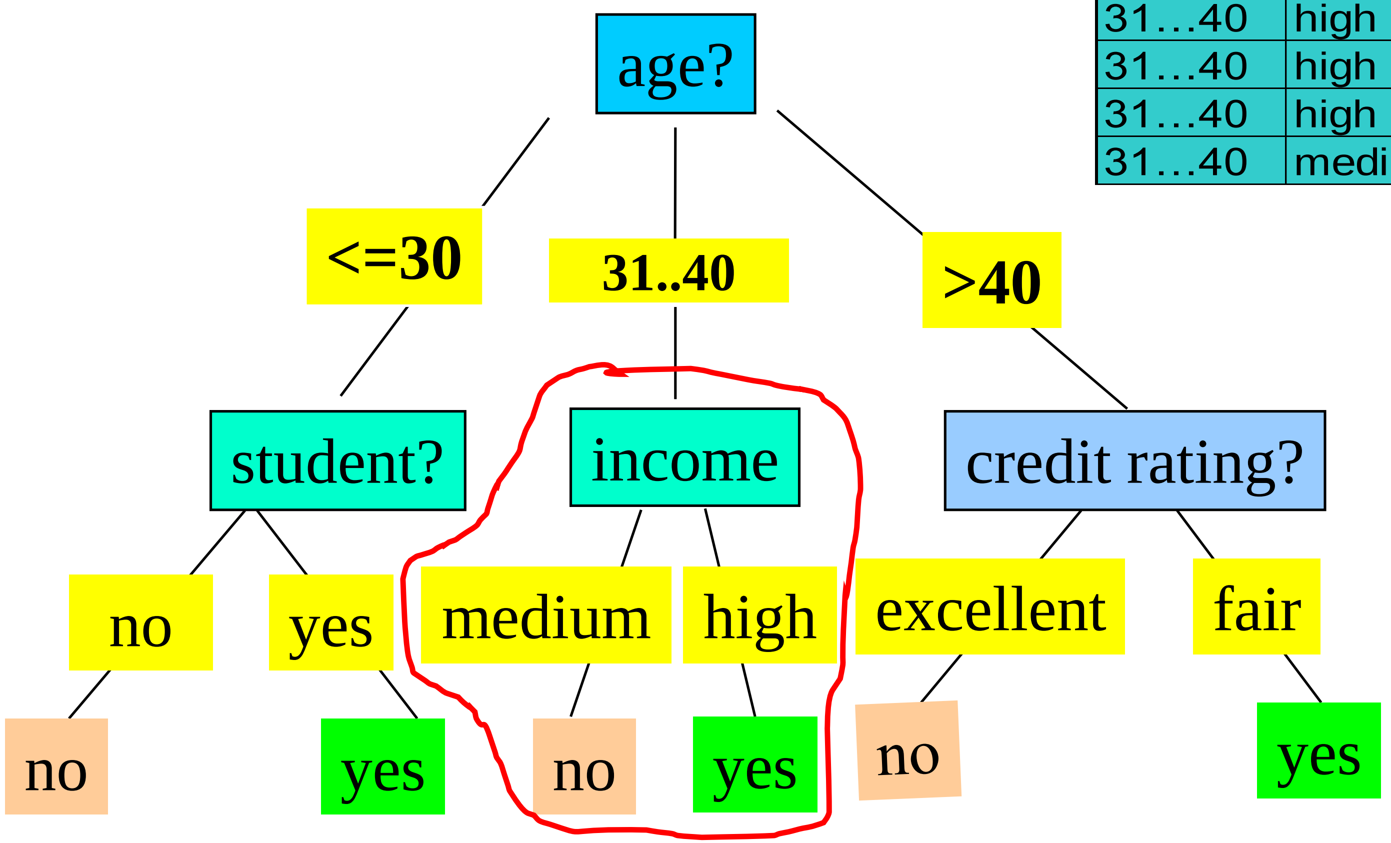
age	income	class
31…40	high	yes
31…40	high	yes
31…40	high	yes
31…40	medium	yes



Overfitting and Tree Pruning

- Overfitting: An induced tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Poor accuracy for unseen samples

age	income	class
31…40	high	yes
31…40	high	yes
31…40	high	yes
31…40	medium	no



Overfitting and Tree Pruning

- Two approaches to avoid overfitting
  - Prepruning: *Halt tree construction early*—do not split a node if this would result in the goodness measure falling below a threshold
    - Difficult to choose an appropriate threshold
  - Postpruning: *Remove branches from a "fully grown" tree*—get a sequence of progressively pruned trees
    - Use a set of data different from the training data to decide which is the "best pruned tree"