08.01.2019

Statistical Methods in AI (CSE/ECE 471)

Lecture-3: Decision Tree Learning

Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad

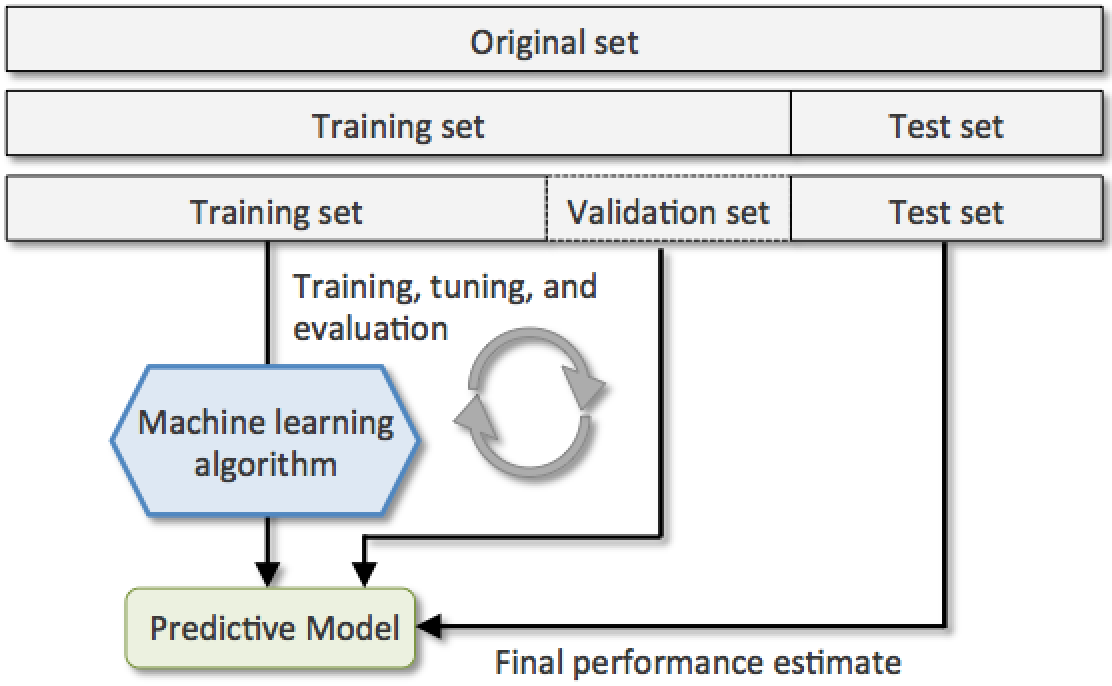
Announcements

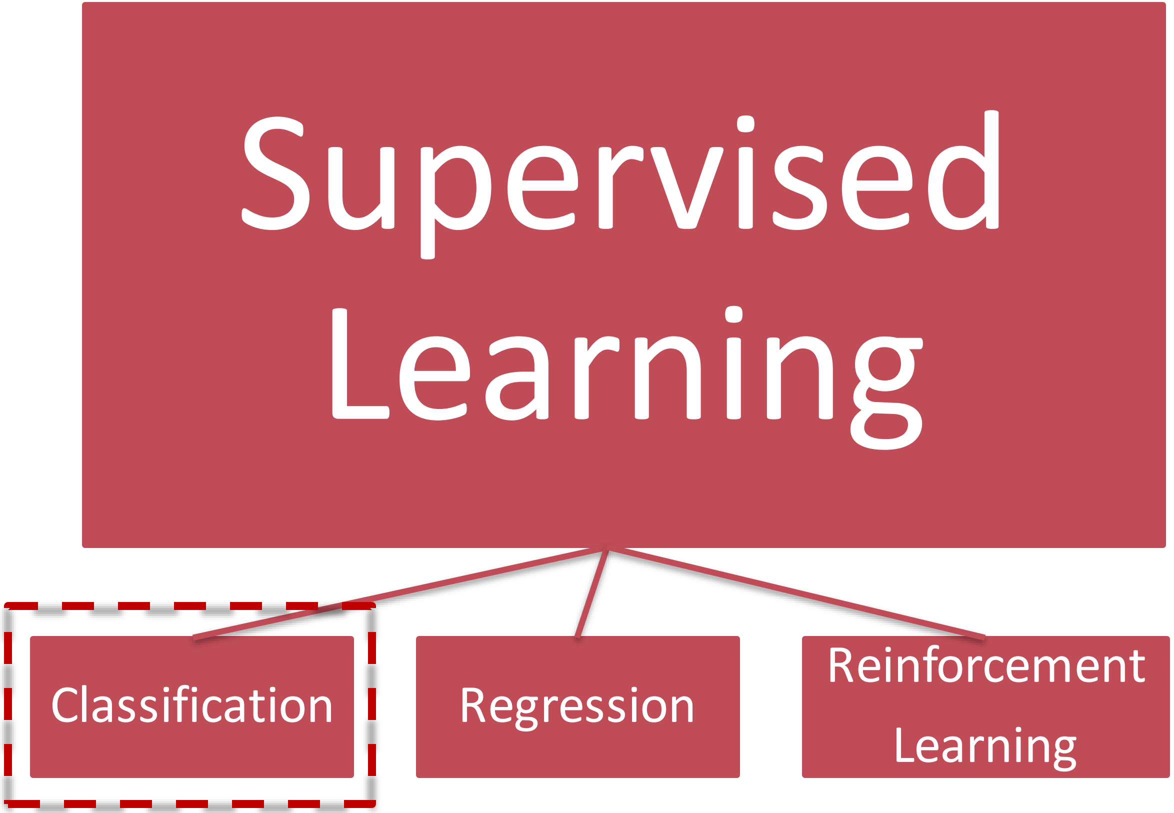
- IMPORTANT: All assignments/projects will need to be in Python.
- This week's tutorial: Probability recap, ML datasets, visualization approaches. **Bring your laptops**.
- Ask questions. Take notes.

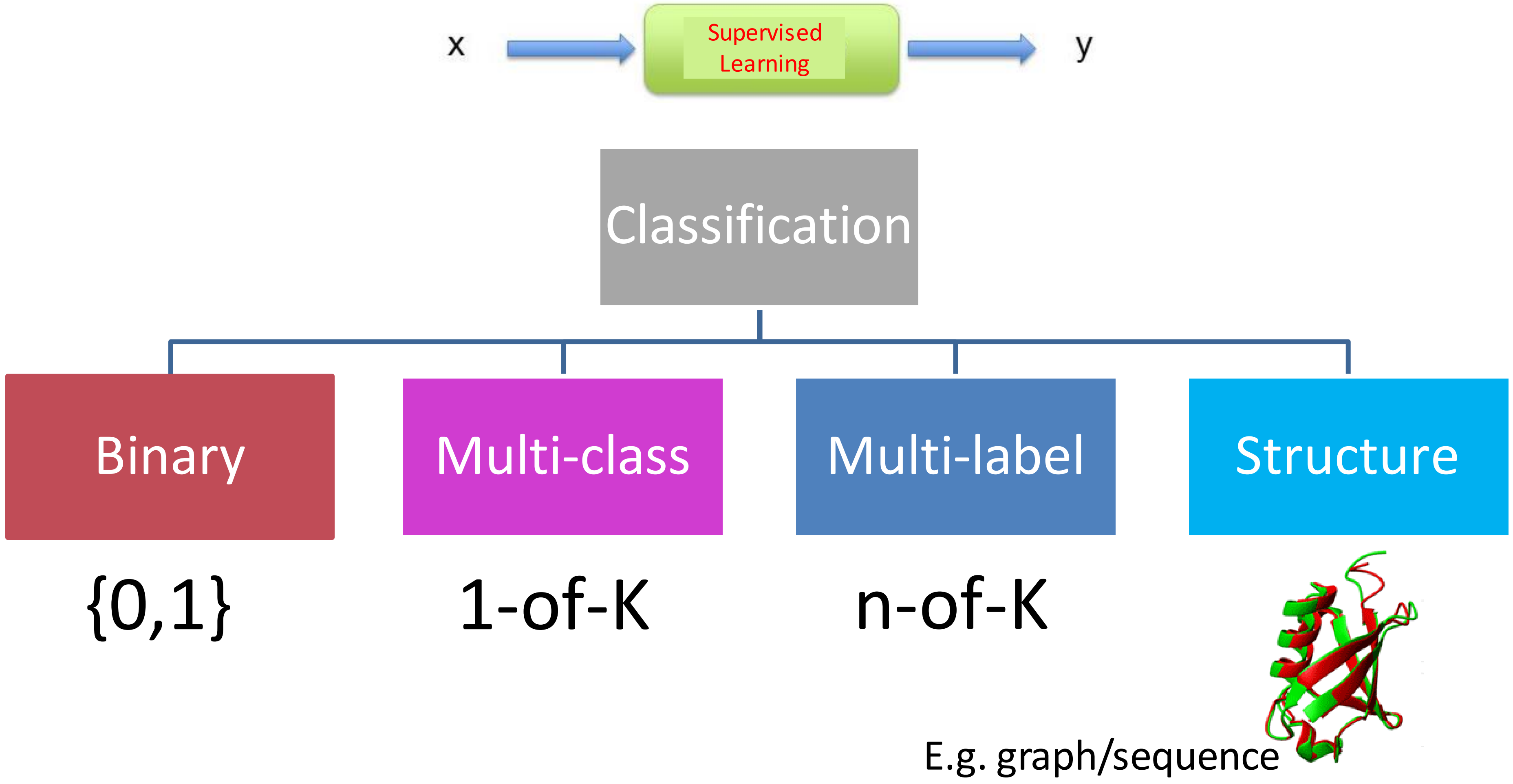
Announcements

- Assignments – Questions involving equations/ mathematical derivation
  - Write up in latex [overleaf.com] → submit
  - Write neatly on paper → scan as photo/pdf [camscanner] → submit
- TA office hours, locations have been announced.

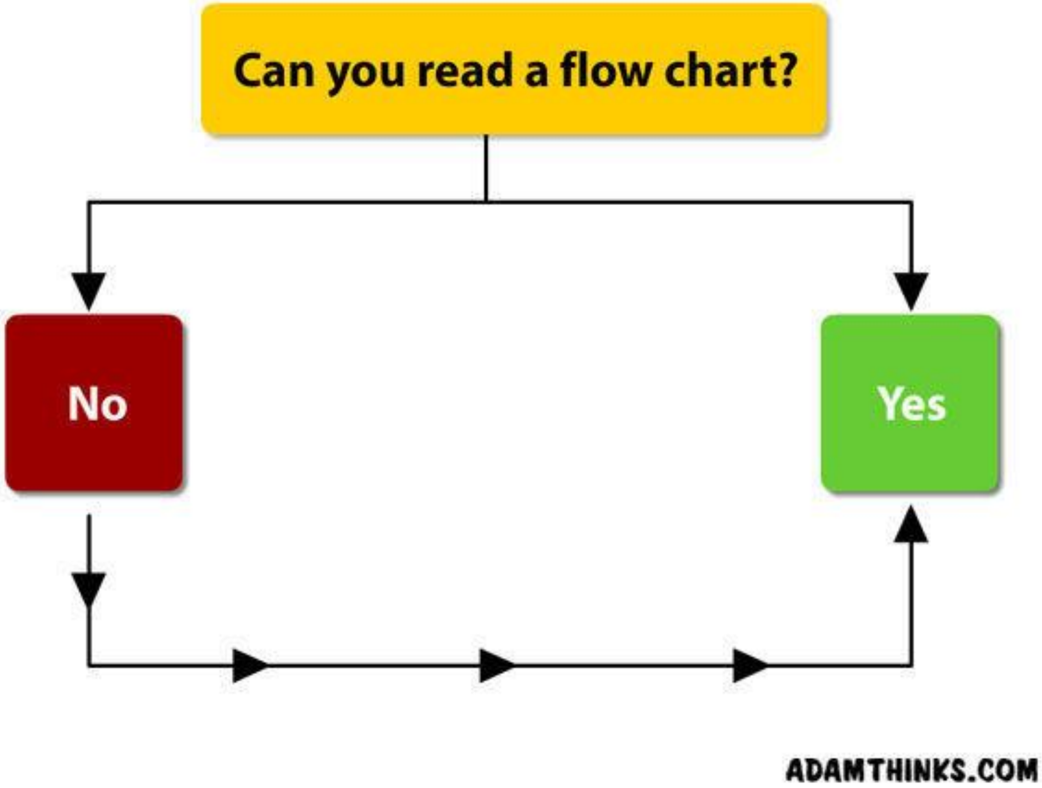
The Train-Validation-Test paradigm




Supervised Learning
Classification
Regression
Reinforcement Learning

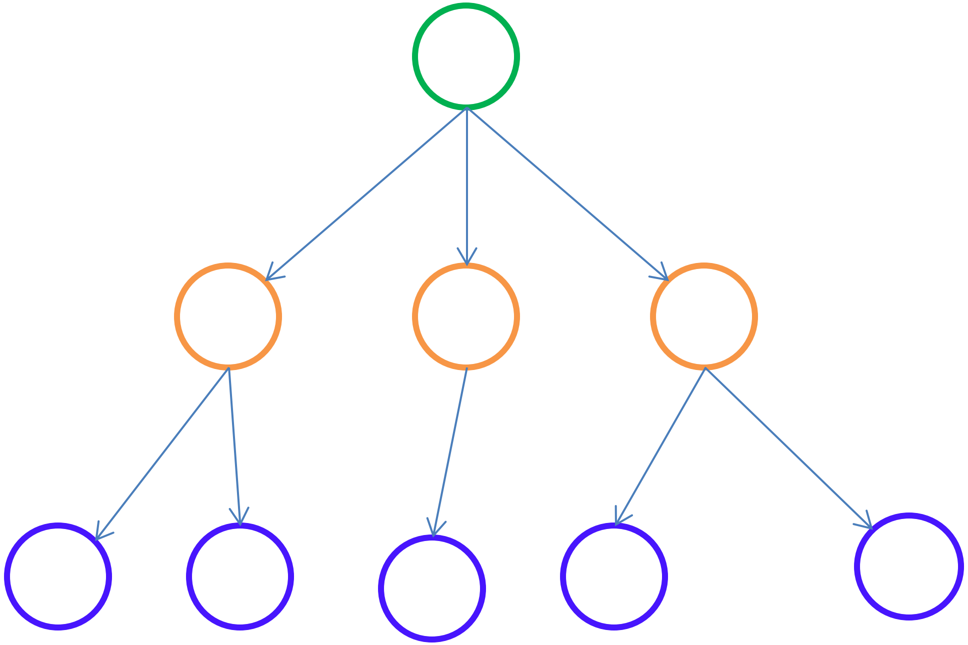

x
Supervised Learning
y
Classification
Binary
Multi-class
Multi-label
Structure
{0,1}
1-of-K
n-of-K
E.g. graph/sequence

Flowcharts

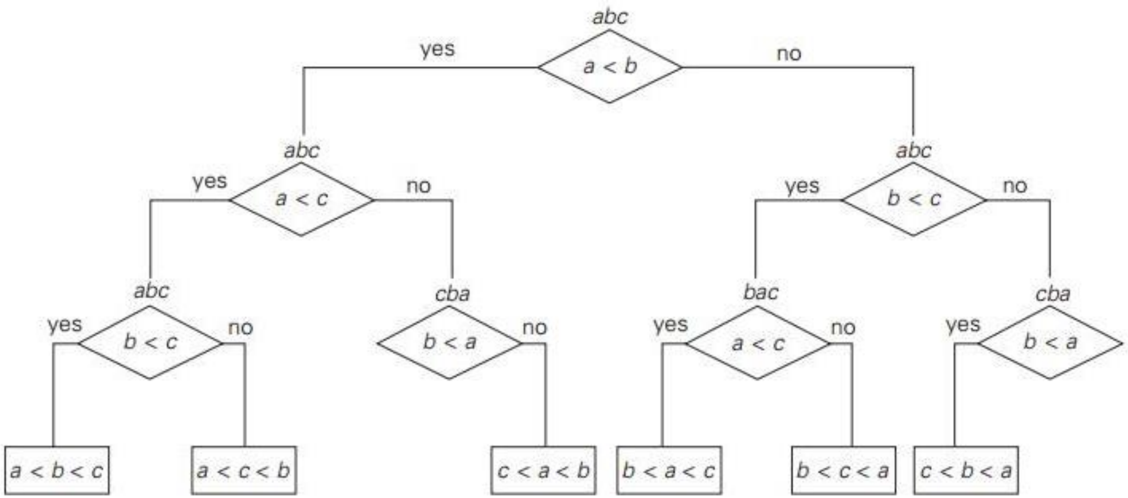


Trees

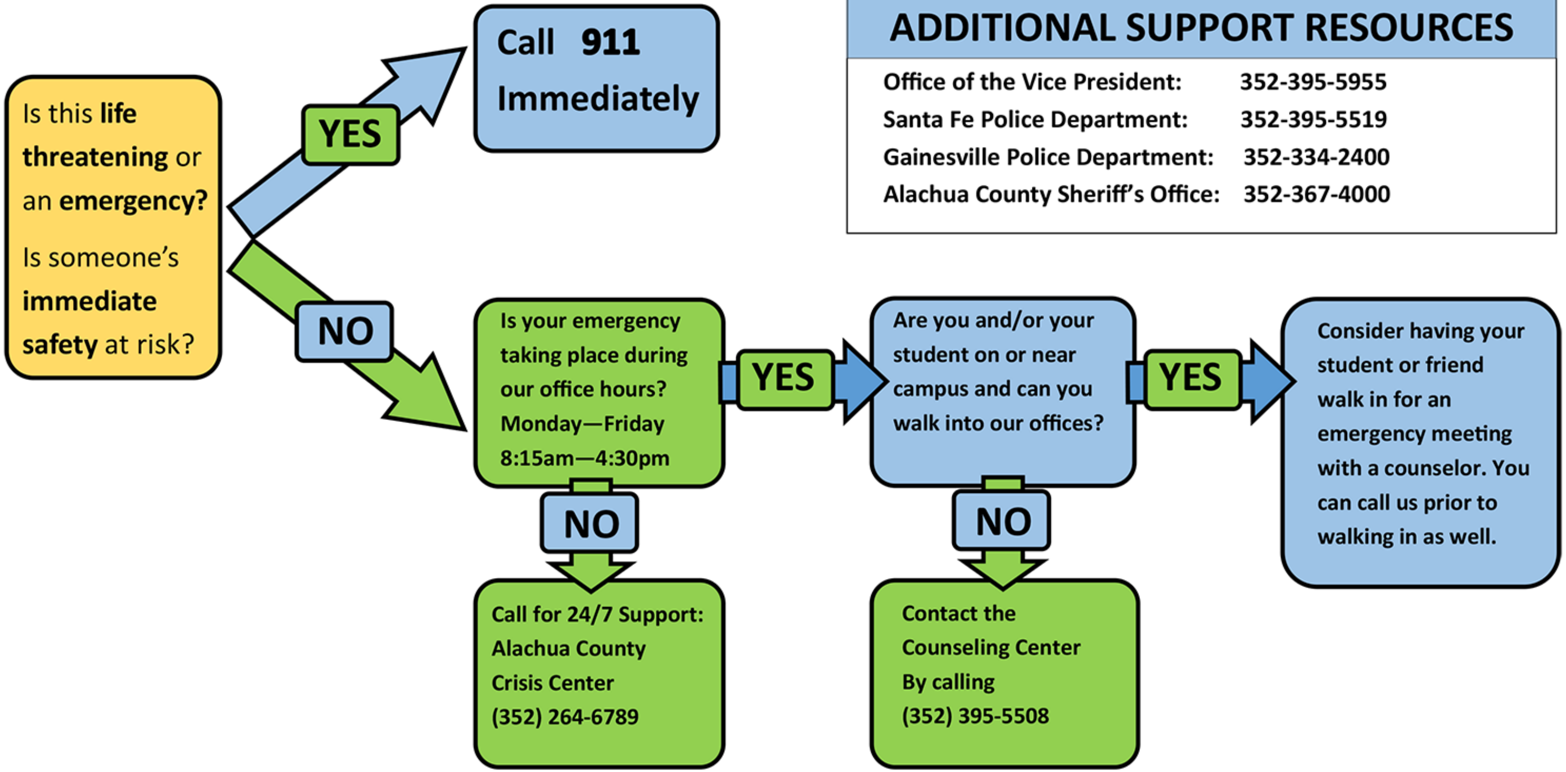
- ❑Node
- ❑Root
- ❑Leaf
- ❑Edge/Branch
- ❑Path
- ❑Depth



if/then
if/else/then

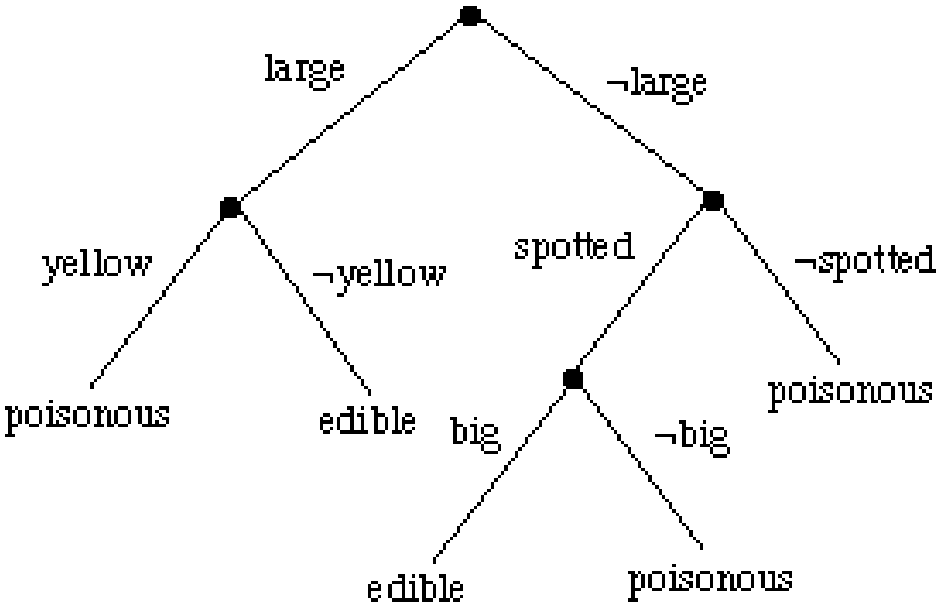


Emergency Response

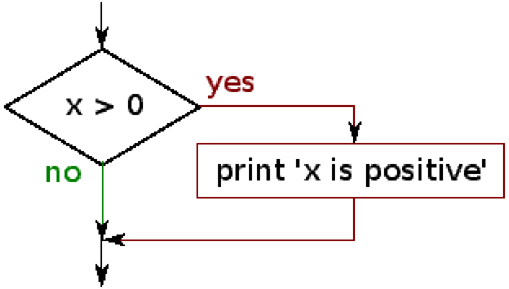


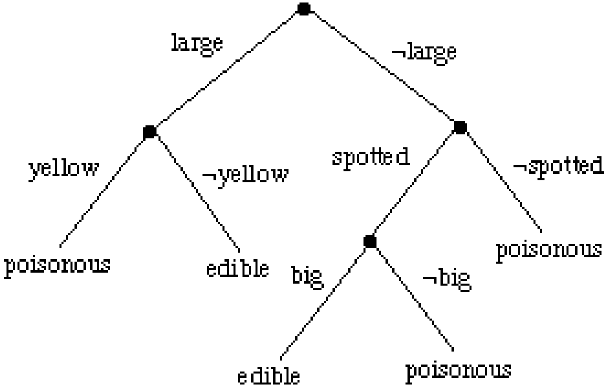
ADDITIONAL SUPPORT RESOURCES	
Office of the Vice President:	352-395-5955
Santa Fe Police Department:	352-395-5519
Gainesville Police Department:	352-334-2400
Alachua County Sheriff's Office:	352-367-4000

Edible Mushroom

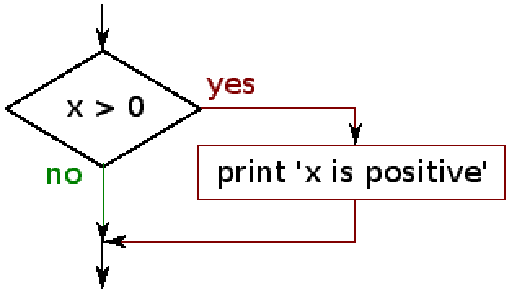


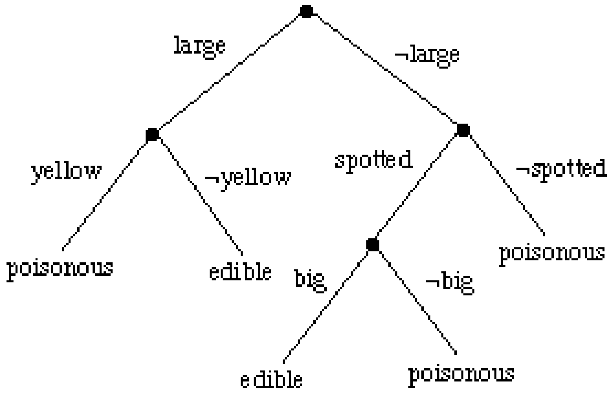
Hand-crafted, fixed trees



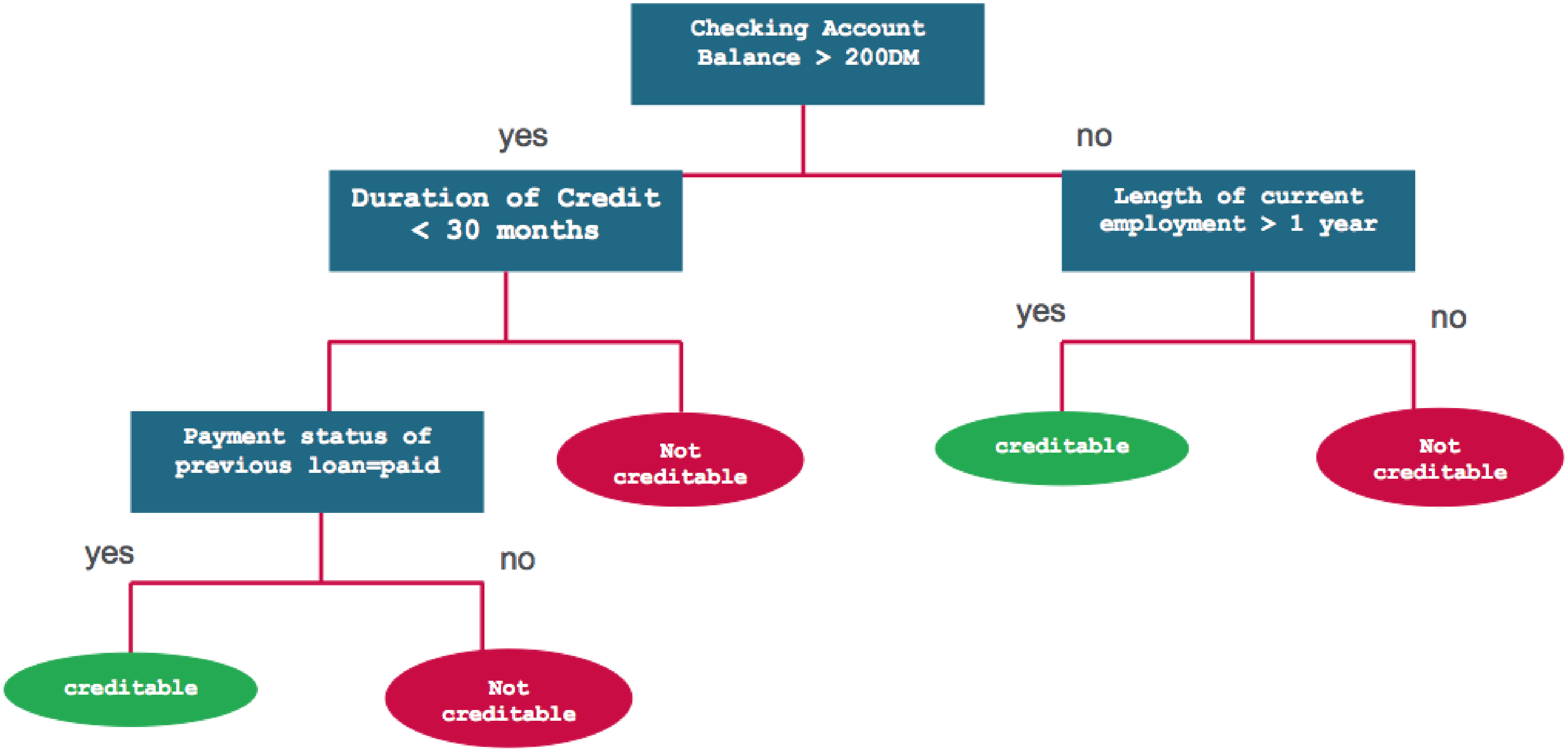


Hand-crafted, fixed trees





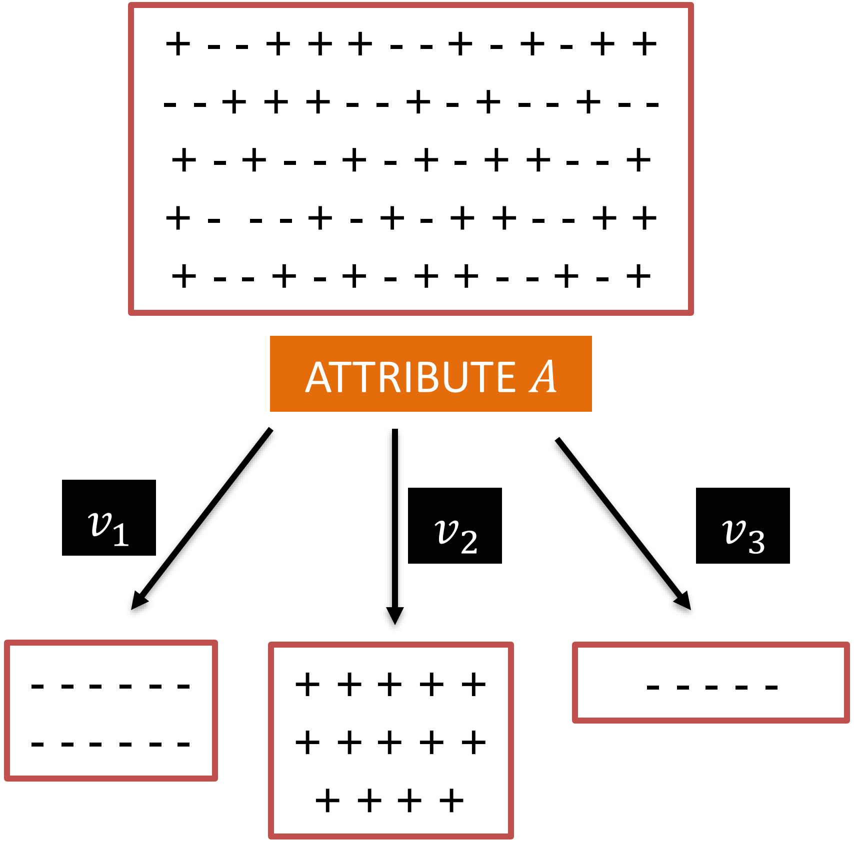
Credit Approval

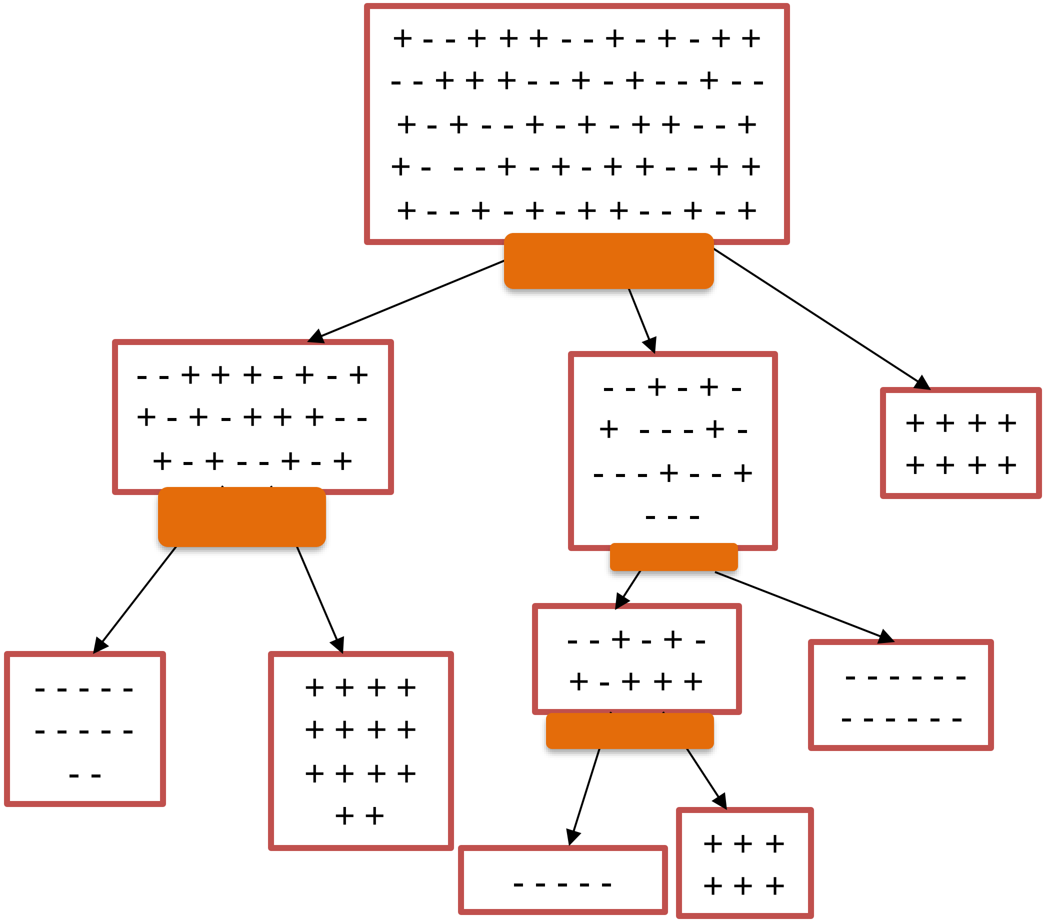


Credit Approval (Raw Data)

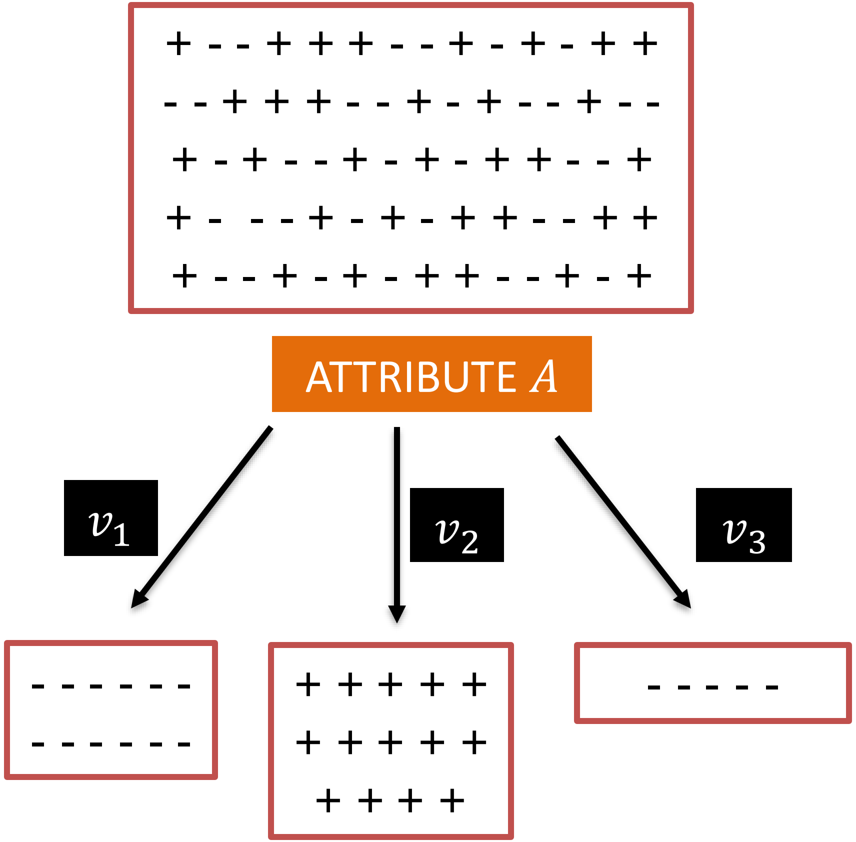
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
64	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
65	a	20.42	0.835	u	g	q	v	1.585	t	t	1	f	g	0	0	+
66	b	26.67	4.25	u	g	cc	v	4.29	t	t	1	t	g	120	0	+
67	b	34.17	1.54	u	g	cc	v	1.54	t	t	1	t	g	520	50000	+
68	a	36	1	u	g	c	v	2	t	t	11	f	g	0	456	+
69	b	25.5	0.375	u	g	m	v	0.25	t	t	3	f	g	260	15108	+
70	b	19.42	6.5	u	g	w	h	1.46	t	t	7	f	g	80	2954	+
71	b	35.17	25.125	u	g	x	h	1.625	t	t	1	t	g	515	500	+
72	b	32.33	7.5	u	g	e	bb	1.585	t	f	0	t	s	420	0	-
73	b	34.83	4	u	g	d	bb	12.5	t	f	0	t	g		0	-
74	a	38.58	5	u	g	cc	v	13.5	t	f	0	t	g	980	0	-
75	b	44.25	0.5	u	g	m	v	10.75	t	f	0	f	s	400	0	-
76	b	44.83	7	y	p	c	v	1.625	f	f	0	f	g	160	2	-
77	b	20.67	5.29	u	g	q	v	0.375	t	t	1	f	g	160	0	-
78	b	34.08	6.5	u	g	aa	v	0.125	t	f	0	t	g	443	0	-

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

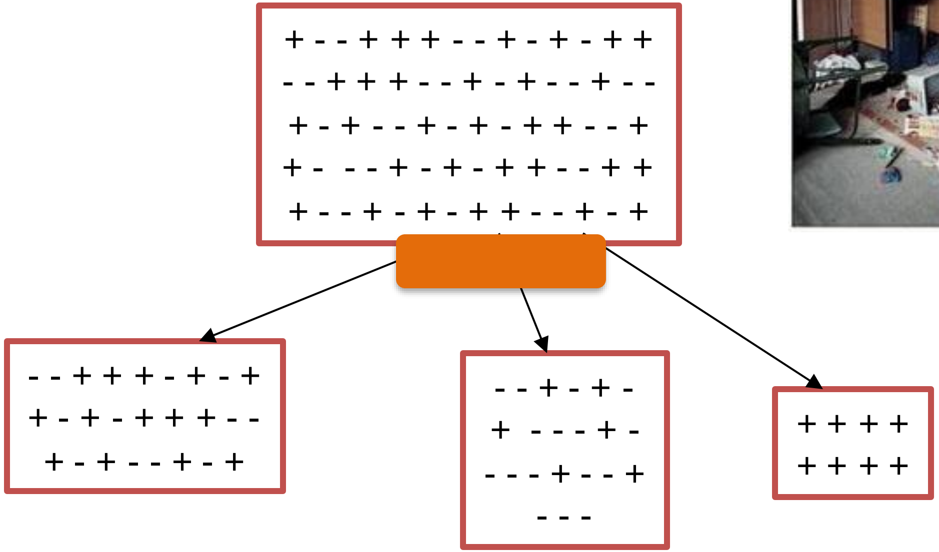

+ - - + + + - - + - + - + +
- - + + + - - + - + - - + - -
+ - + - - + - + - + + - - +
+ - - - + - + - + + - - + +
+ - - + - + - + + - - + - +
ATTRIBUTE A
v_1
v_2
v_3
- - - - - -
- - - - - -
+ + + + +
+ + + + +
+ + + +
- - - - -



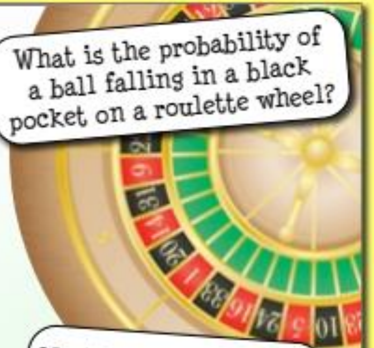
How much ‘impurity’ does this attribute decrease ?



How much ‘impurity’ does this attribute decrease ?



PROBABILITY



What is the probability of a ball falling in a black pocket on a roulette wheel?

Probability is the chance (or likelihood) of an event happening. We can describe probability using words from a probability scale:

← less likely | more likely →

no chance | poor chance | even chance | good chance | certain

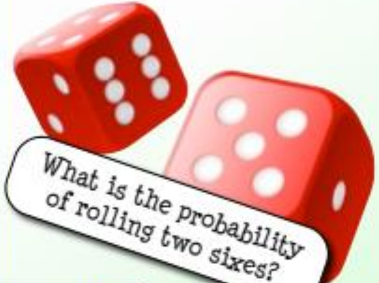
We can also describe probability using fractions.

The probability of a flipped coin landing on heads is one out of two. $\frac{1}{2}$

The probability of this spinner landing on '3' is three out of six. $\frac{3}{6}$

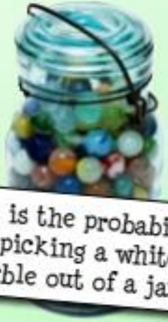
The probability of rolling a '2' on a die is one out of six. $\frac{1}{6}$

What does 'random' mean?
If you put your hand into a bag of marbles and took one out without looking, you picked it at **random**. You didn't choose that one - you got it by chance.

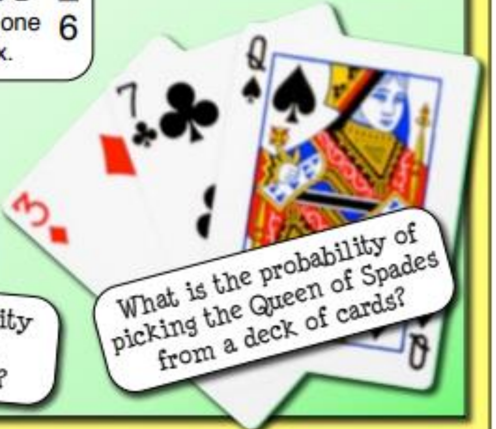


What is the probability of rolling two sixes?

What is the probability of a flipped coin landing on tails?



What is the probability of picking a white marble out of a jar?



What is the probability of picking the Queen of Spades from a deck of cards?

Discrete probability distribution

- Variables that have only a finite number of possible outcomes
- For example …a six-sided die is thrown

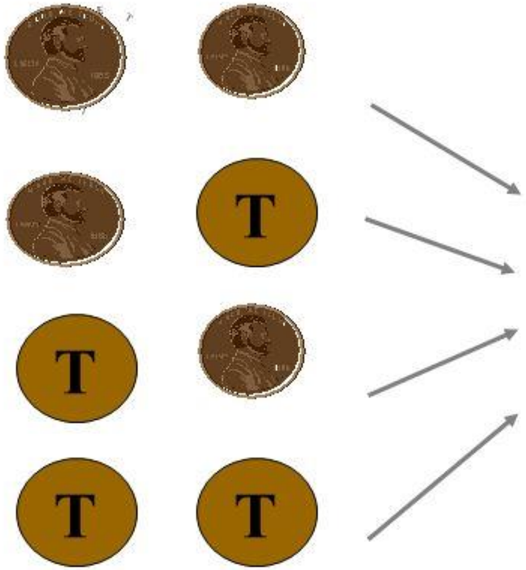
Possibilities $r=1$	1	2	3	4	5	6
Probability that $Z=r$	1/6	1/6	1/6	1/6	1/6	1/6

- $$0 \leq P\left(X_j\right) \leq 1 \qquad \sum P\left(X_j\right) = 1$$

Discrete probability distribution

Event: Toss two coins

Count the number of tails

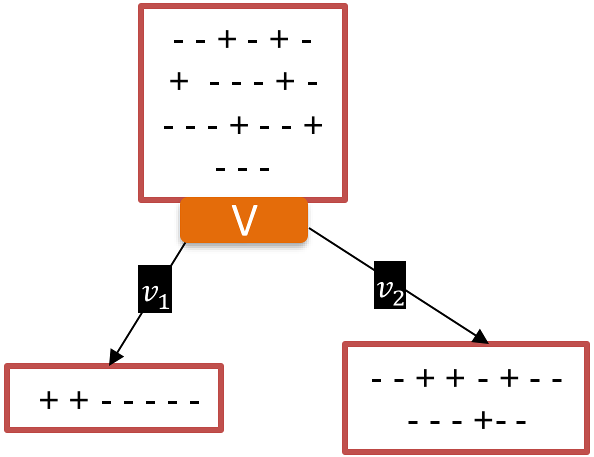


Probability Distribution

Values	Probability
0	1/4 = .25
1	2/4 = .50
2	1/4 = .25

Properties of an impurity measure

- Class labels: Binary {+1,-1}
- q



An ***impurity measure*** is a function $i(V)$ such that

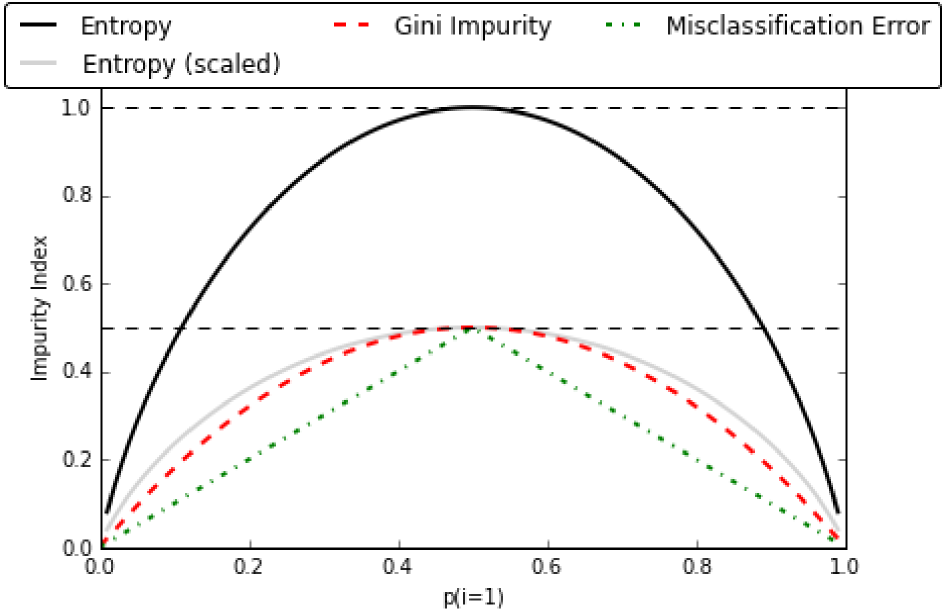
- $i(V) \geq 0$, with $i(V) = 0$ iff V consists of a single class
- a larger value of $i(V)$ indicates that the distribution defined by $(q, (1-q))$ is closer to the uniform distribution

Impurity function: candidates

Entropy: $i(V) = -\left(q \log q + (1-q)\log(1-q)\right)$

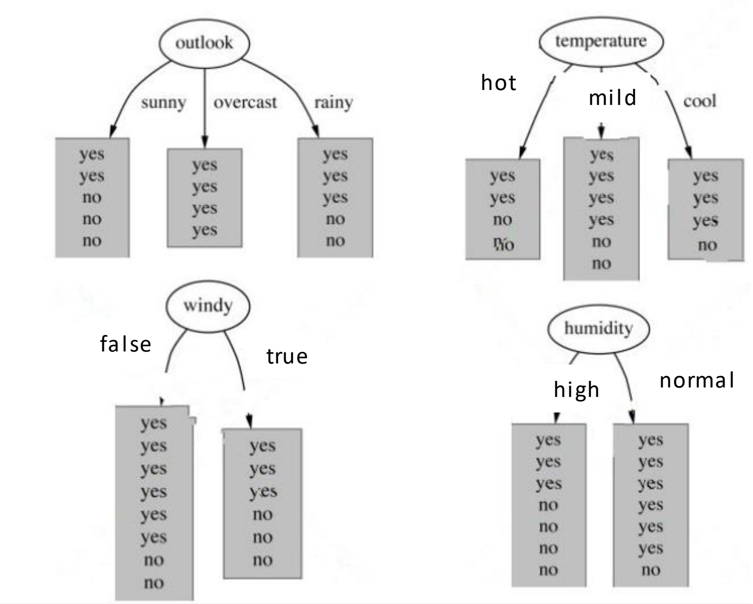
Gini index: $i(V) = 2q(1-q)$

Misclassification rate: $i(V) = \min(q, 1-q)$



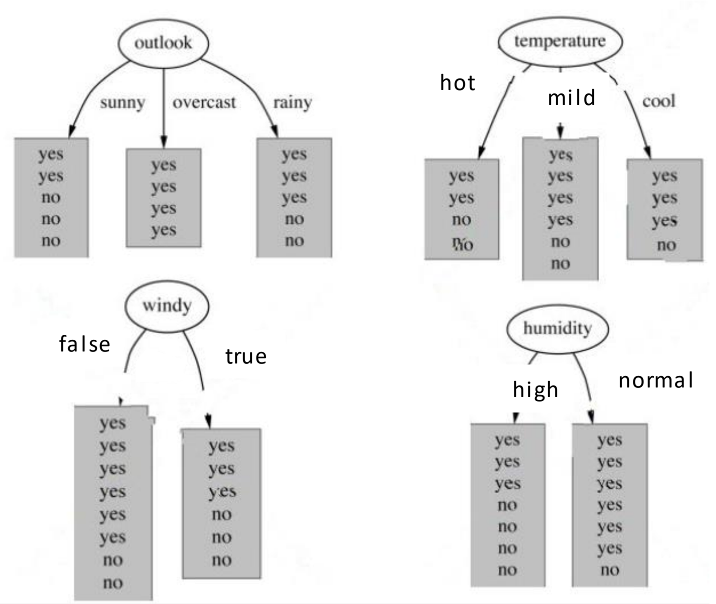
Example

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes



Step-1: Compute impurity score of training label distribution

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes



Entropy: $i(V) = -(q \log q + (1-q) \log(1-q))$

$$E(S) = -(\frac{9}{14}log(\frac{9}{14}) + \frac{5}{14}log(\frac{5}{14})) = 0.94$$

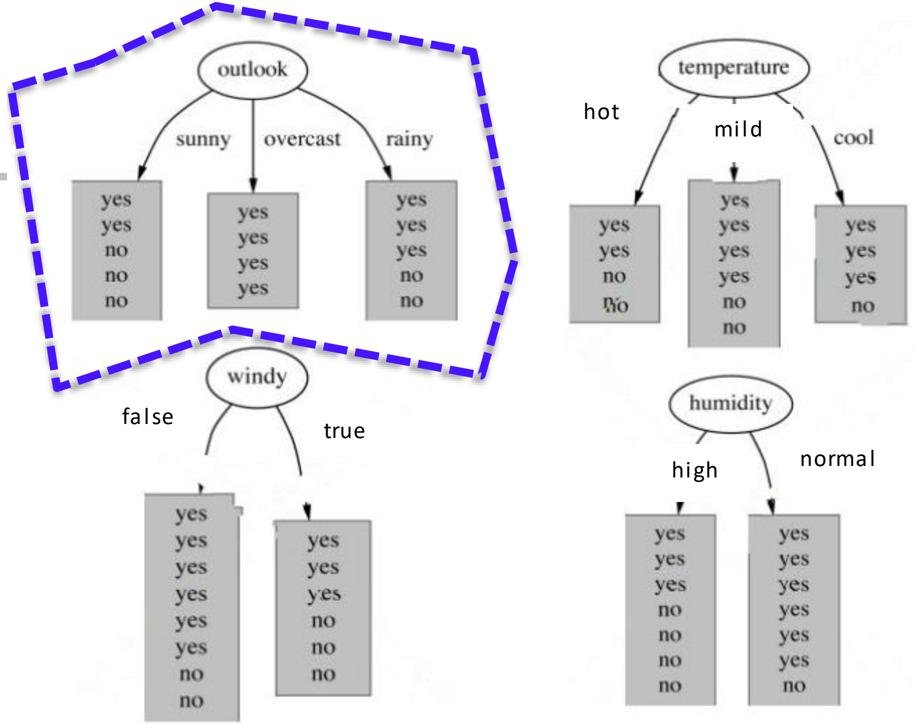
Step-2: Compute impurity score for each unique value of candidate attributes

Example: Attribute Outlook

Entropy: $i(V) = -(q \log q + (1-q)\log(1-q))$

- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$



Step-2: Compute impurity score for each unique value of candidate attributes

Example: Attribute Outlook

Entropy: $i(V) = -(q \log q + (1-q) \log(1-q))$

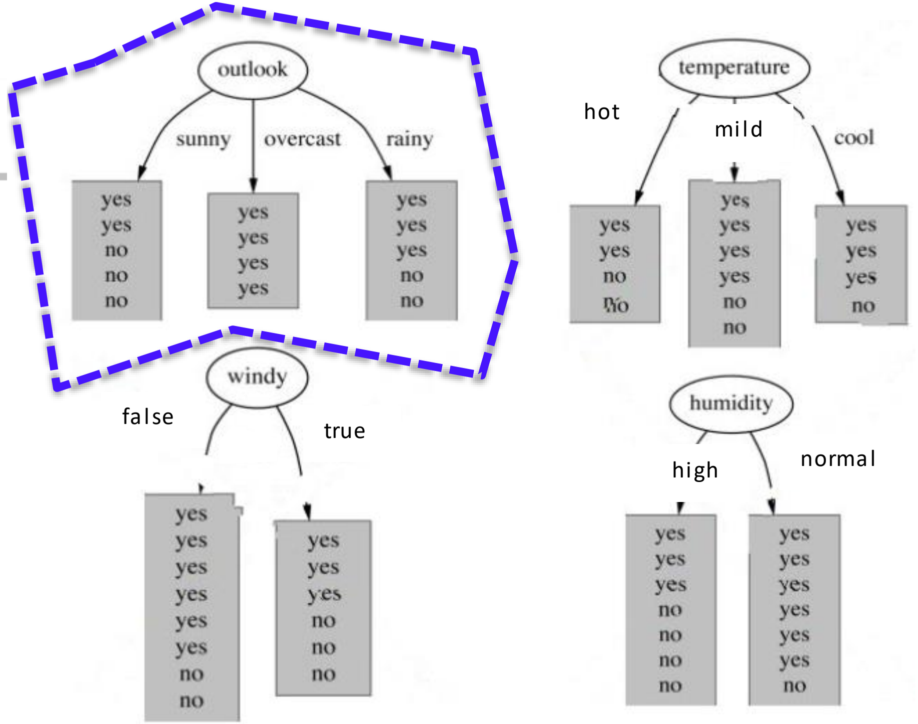
- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook} = \text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

- Outlook = overcast: 4 examples yes, 0 examples no

$$E(\text{Outlook} = \text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0



Step-2: Compute impurity score for each unique value of candidate attributes

Example: Attribute Outlook

Entropy: $i(V) = -(q \log q + (1-q)\log(1-q))$

- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

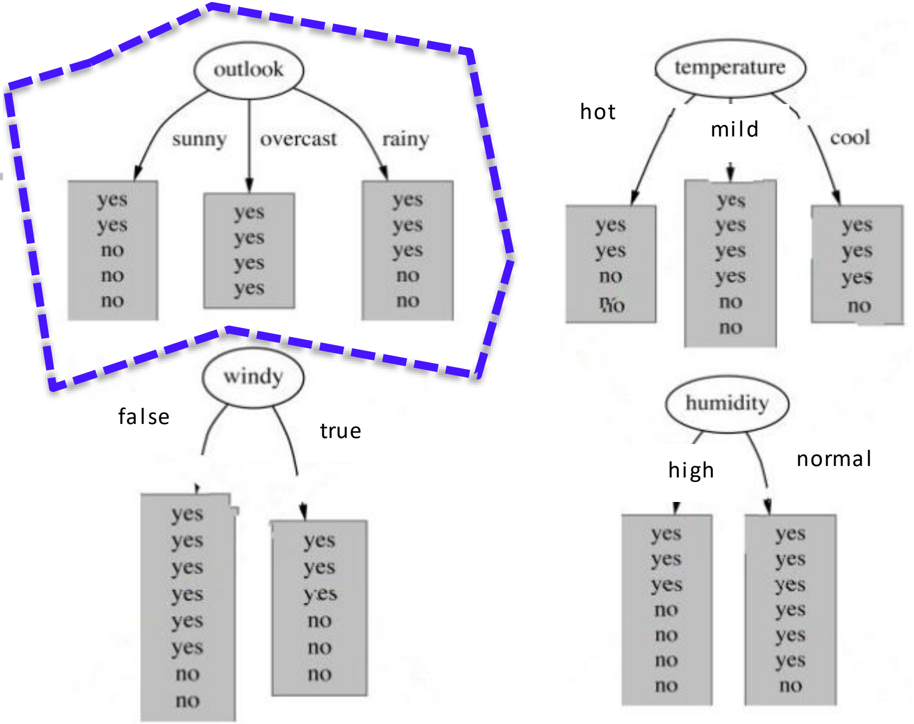
- **Outlook = overcast:** 4 examples yes, 0 examples no

$$E(\text{Outlook}=\text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0

- **Outlook = sunny** 2 examples yes, 3 examples no

$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$



Step-3: Compute impurity score for candidate attribute

- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

- Outlook = overcast: 4 examples yes, 0 examples no

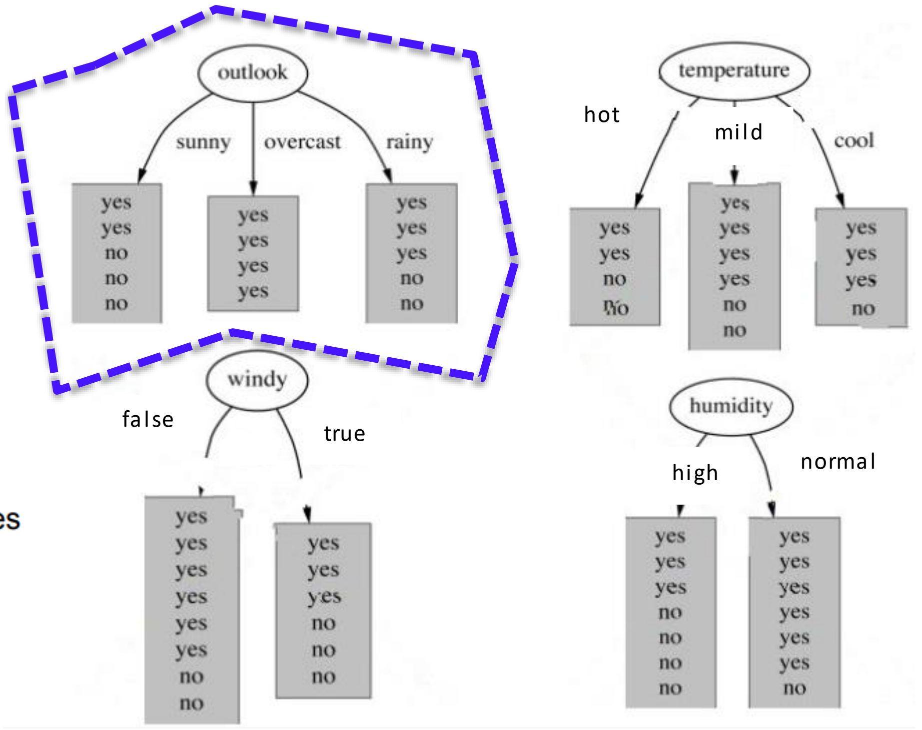
$$E(\text{Outlook}=\text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0

- **Outlook = sunny** 2 examples yes, 3 examples no

$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$

- Entropy only computes the quality of a single (sub-)set of examples
  - corresponds to a single value
- How can we compute the quality of the entire split?
  - corresponds to an entire attribute



Step-3: Compute impurity score for candidate attribute

- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

- Outlook = overcast: 4 examples yes, 0 examples no

$$E(\text{Outlook}=\text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0

- **Outlook = sunny** 2 examples yes, 3 examples no

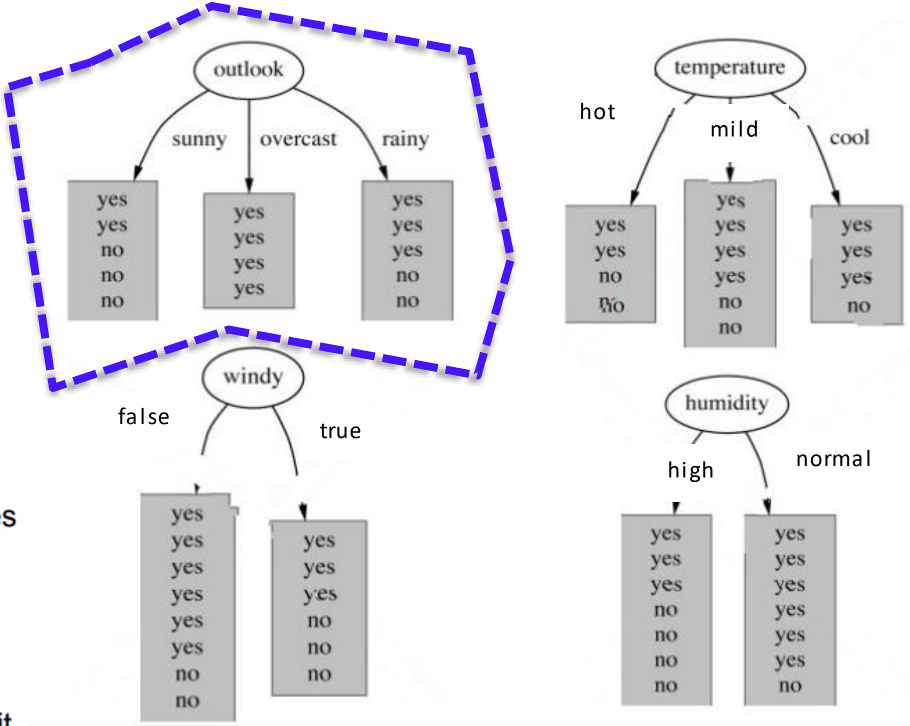
$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$

- Entropy only computes the quality of a single (sub-)set of examples
  - corresponds to a single value
- How can we compute the quality of the entire split?
  - corresponds to an entire attribute

Solution:

- Compute the weighted average over all sets resulting from the split
  - weighted by their size

$$I(S, A) = \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$$



Step-3: Compute impurity score for candidate attribute

- **Outlook = rainy** 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

- Outlook = overcast: 4 examples yes, 0 examples no

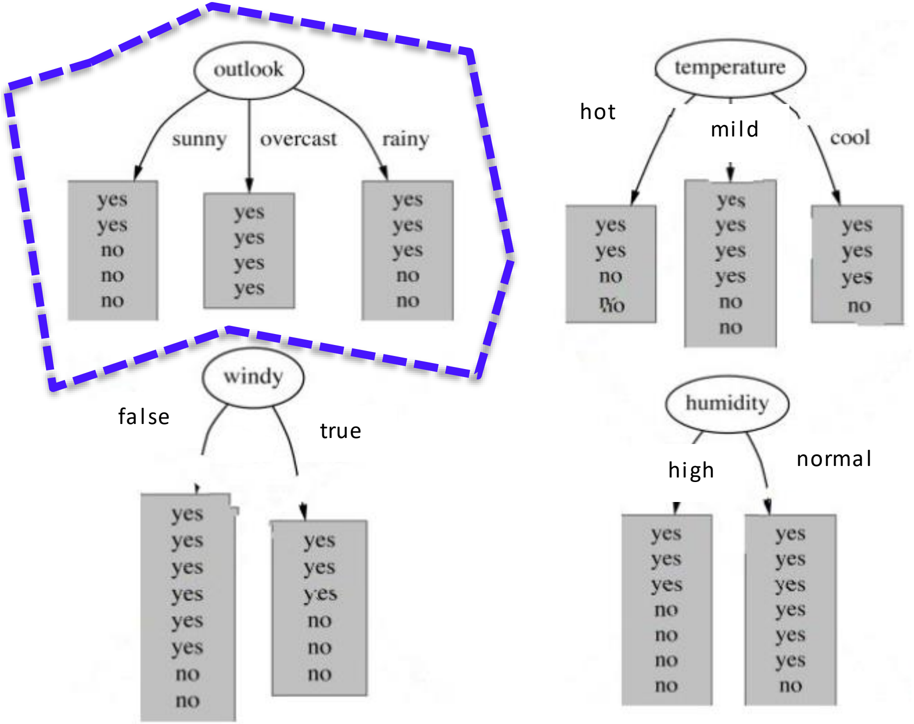
$$E(\text{Outlook}=\text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0

- **Outlook = sunny** 2 examples yes, 3 examples no

$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$

$$I(S,A) = \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$$



- Average entropy for attribute *Outlook*:

$$I(\text{Outlook}) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.693$$

Step-4: Compute Information Gain (reduction in impurity score) provided by candidate attribute

$$I(S,A)=\sum_i \frac{|S_i|}{|S|}\cdot E(S_i)$$

- Average entropy for attribute *Outlook*:

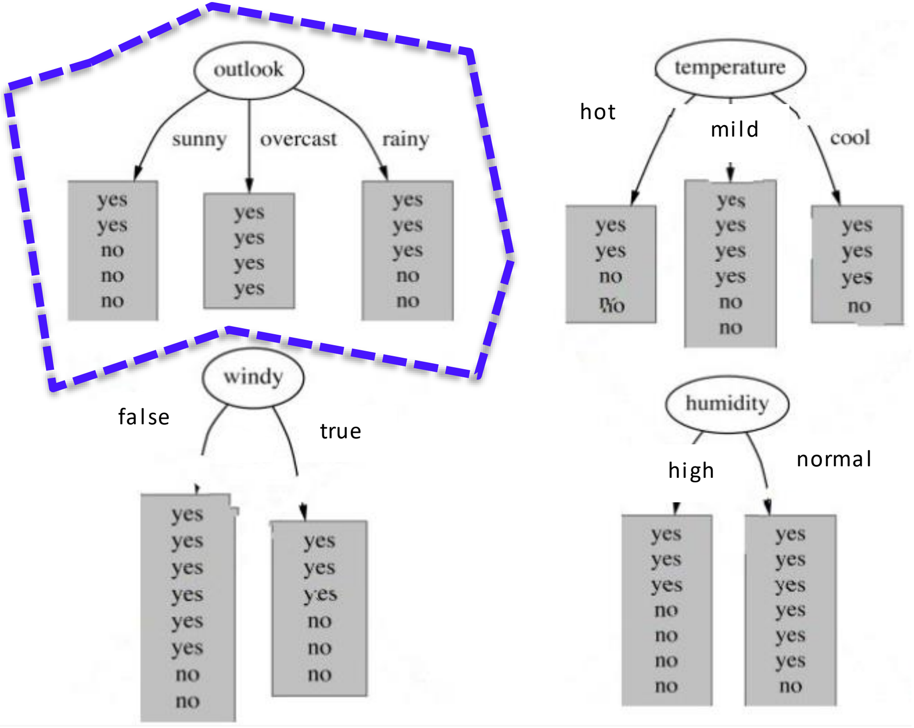
$$I(\text{Outlook})=\frac{5}{14}\cdot 0.971+\frac{4}{14}\cdot 0+\frac{5}{14}\cdot 0.971=0.693$$

$$E(S) = -(\frac{9}{14}log(\frac{9}{14}) + \frac{5}{14}log(\frac{5}{14})) = 0.94$$

Information Gain for Attribute A

$$Gain(S,A)=E(S)-I(S,A)=E(S)-\sum_i \frac{|S_i|}{|S|}\cdot E(S_i)$$

$$Gain(S,Outlook)=0.246$$



Step-5: Compare Information Gain provided by all candidates

Information Gain for Attribute A

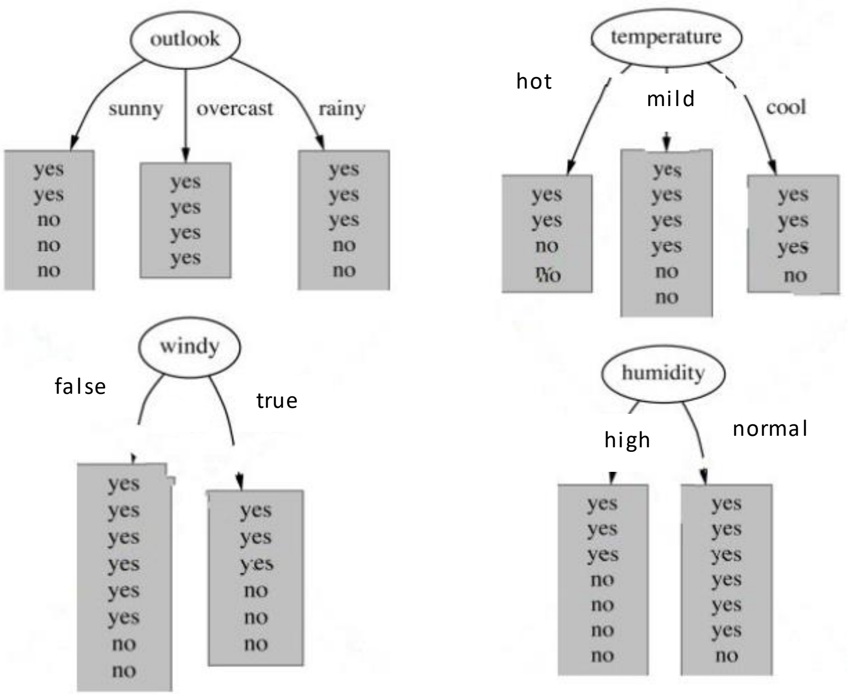
$$Gain(S,A)=E(S)-I(S,A)=E(S)-\sum_i \frac{|S_i|}{|S|}\cdot E(S_i)$$

Gain (S, Humidity)
= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)
= .940 - (8/14).811 - (6/14)1.0
= .048

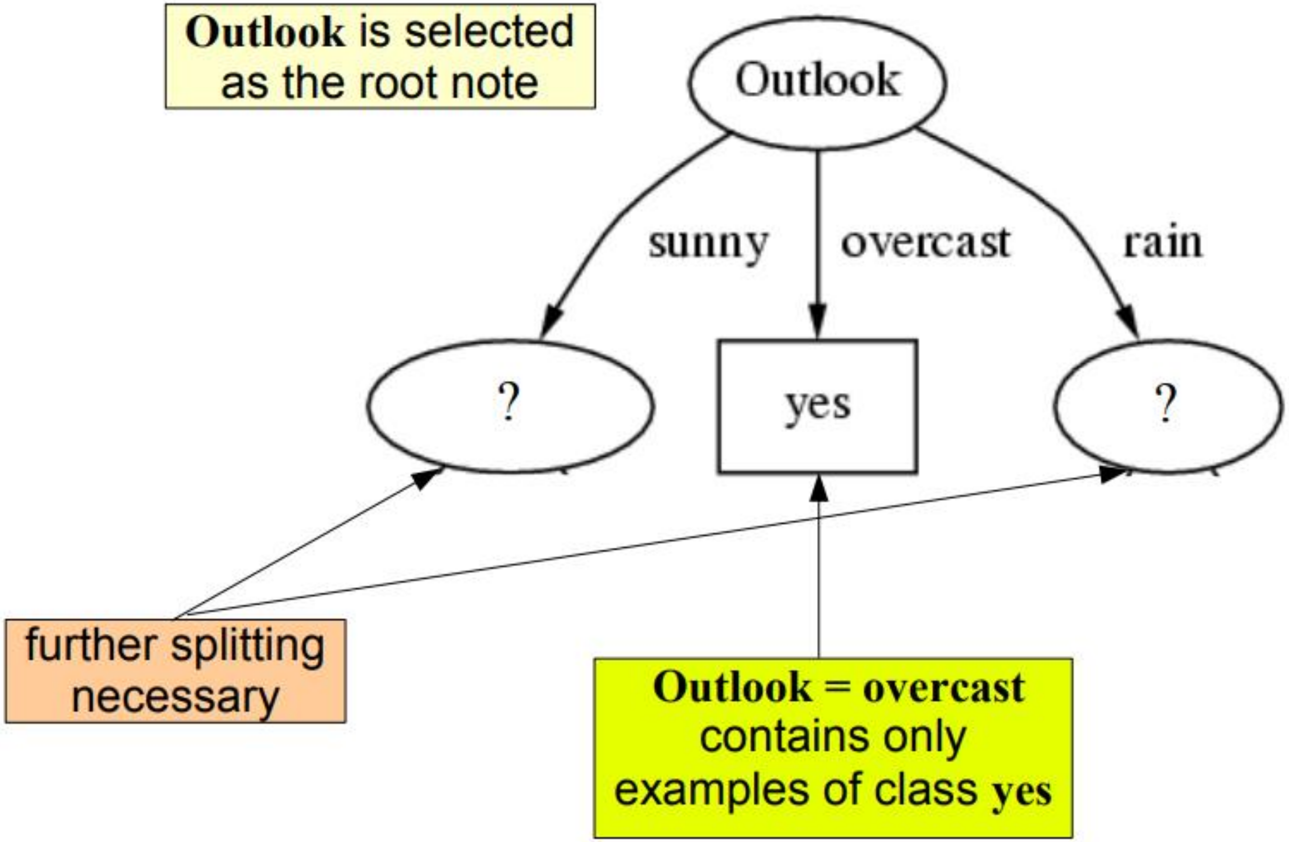
$Gain(S, Outlook) = \boxed{0.246}$

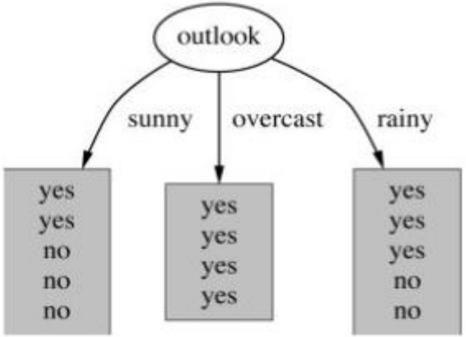
$Gain(S, Temperature) = 0.029$



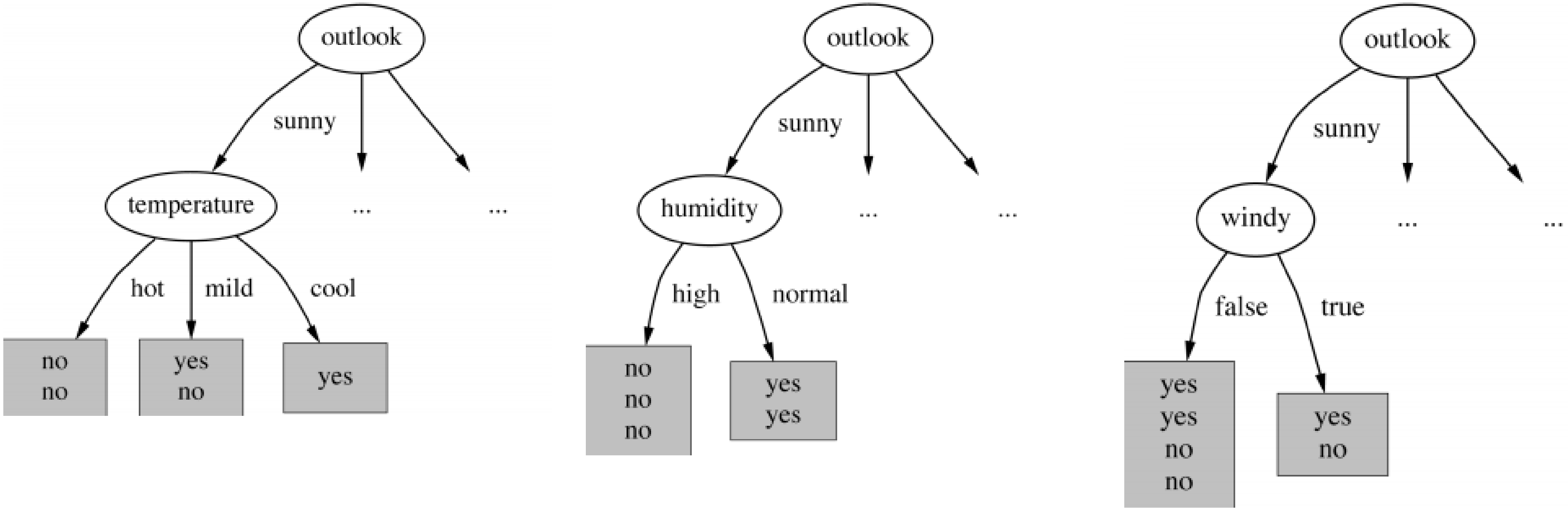
Select attribute which provides largest 'impurity reduction'/Information Gain

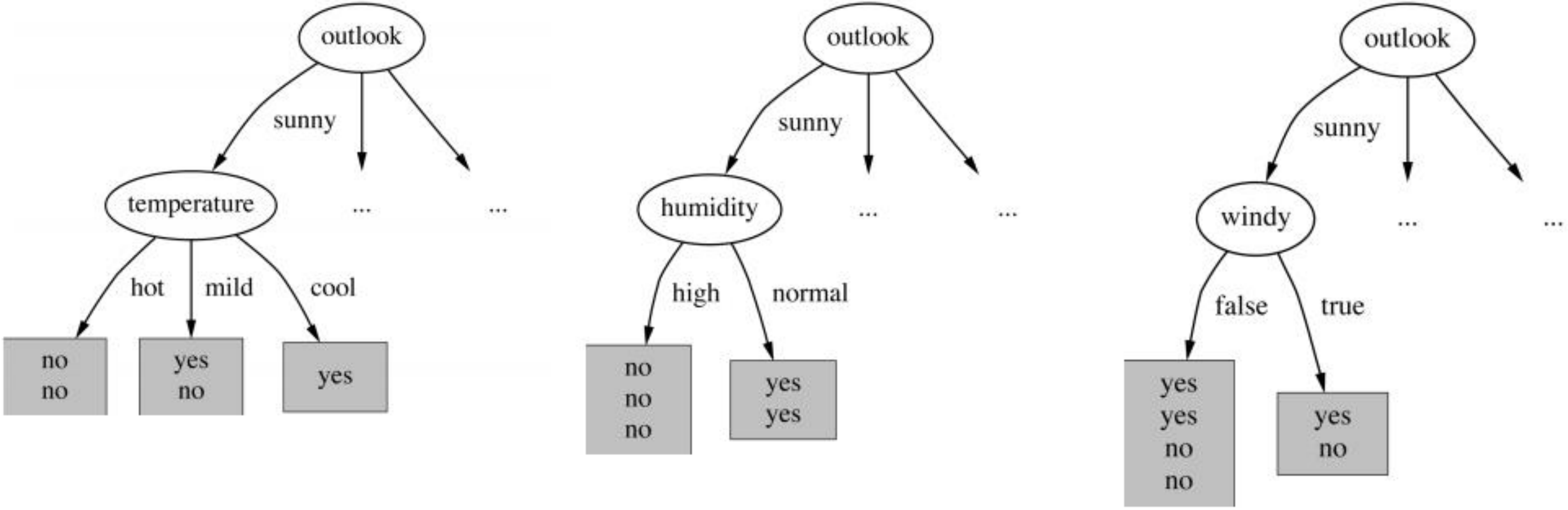
Step-6: Assign root node





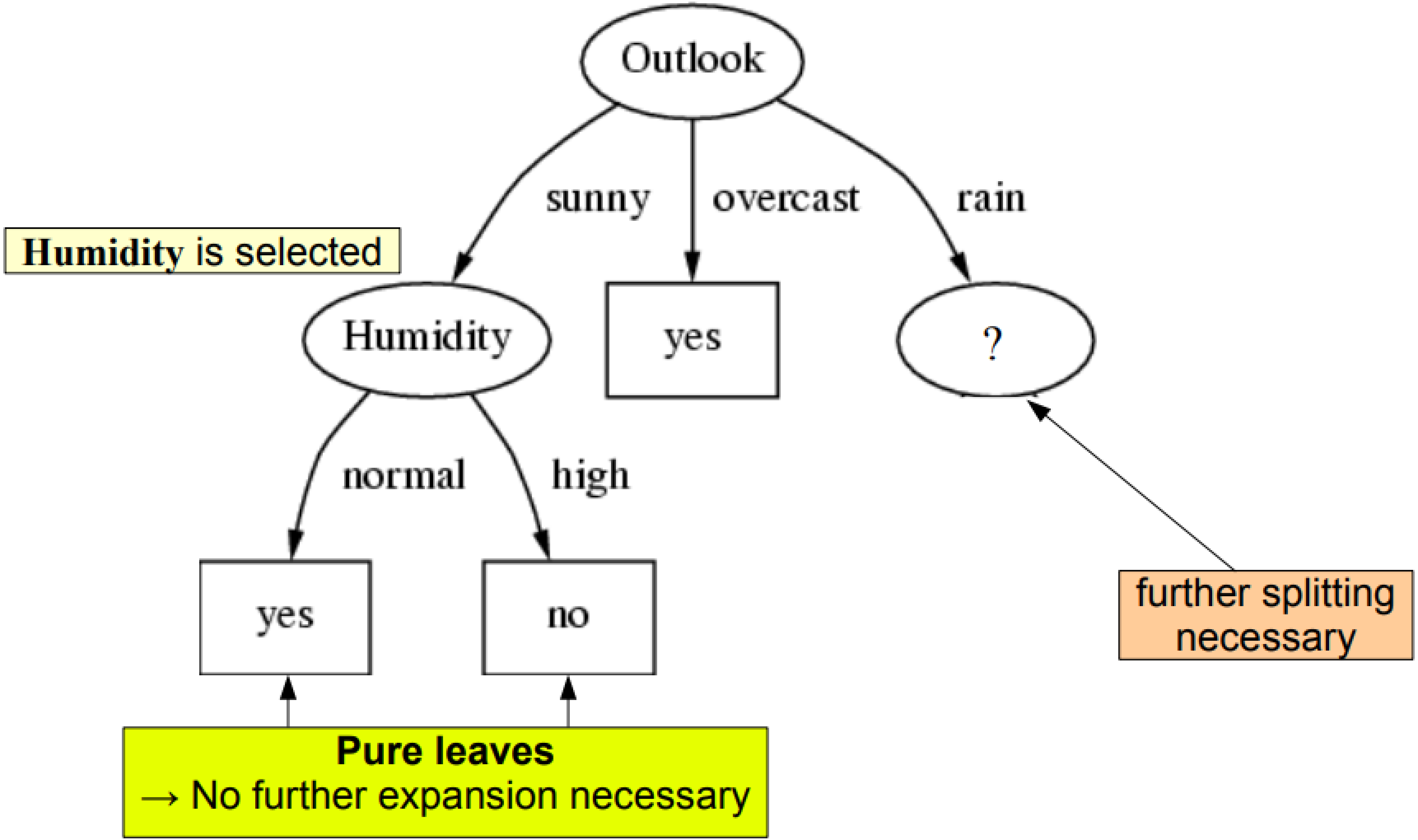
Recurse and repeat Steps 1-6



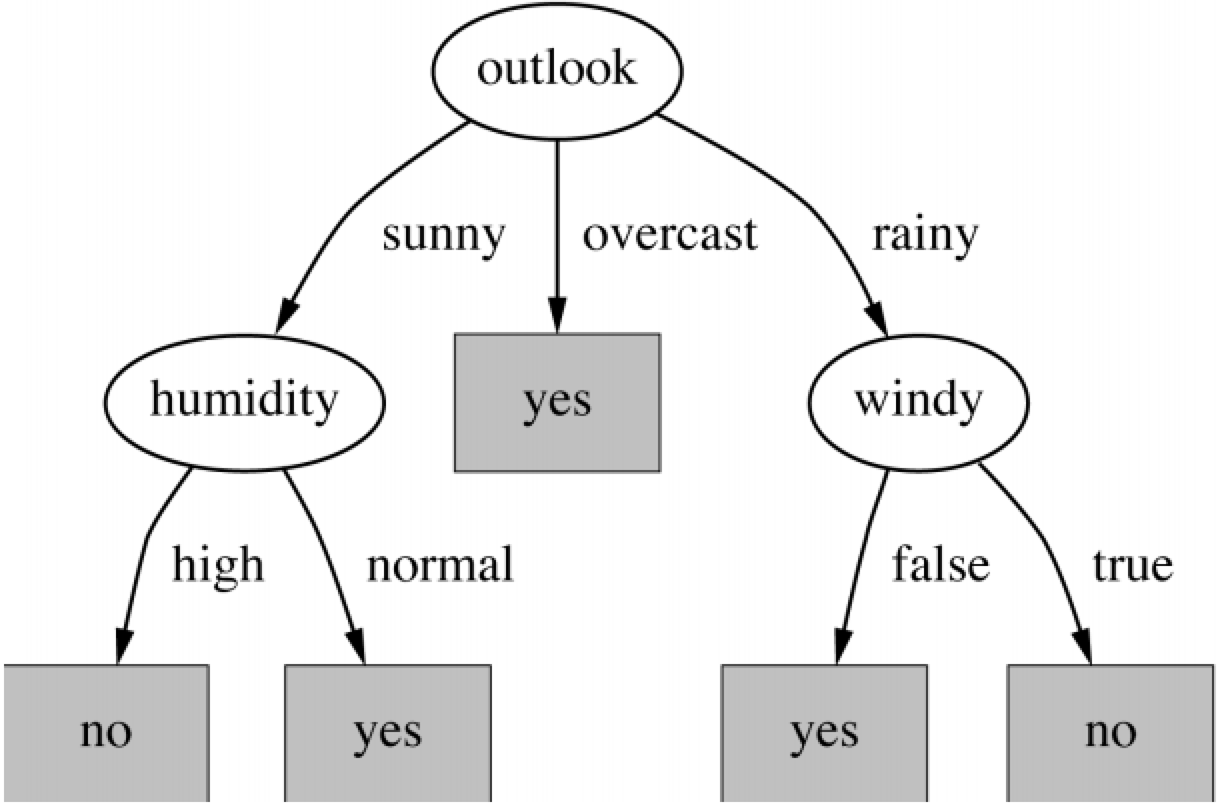


$$
\left.\begin{aligned}
\text{Gain}(\textit{Temperature}\,) &= 0.571 \text{ bits} \\
\text{Gain}(\textit{Humidity}\,) &= 0.971 \text{ bits} \\
\text{Gain}(\textit{Windy}\,) &= 0.020 \text{ bits}
\end{aligned}\right\}
$$

Humidity is selected

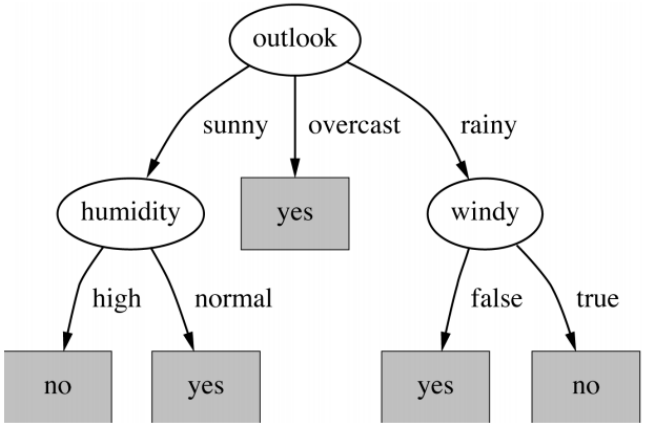

Outlook
sunny
overcast
rain
Humidity is selected
Humidity
yes
?
normal
high
yes
no
further splitting necessary
Pure leaves
→ No further expansion necessary

Final Decision Tree



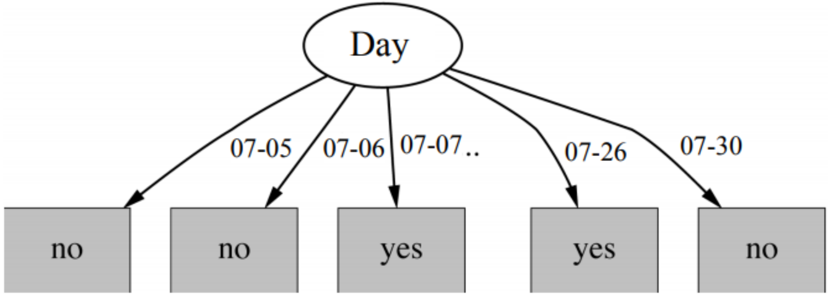
Final Decision Tree

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes



- Problematic: attributes with a large number of values
  - extreme case: each example has its own value
    - e.g. example ID; Day attribute in weather data

- Problematic: attributes with a large number of values
  - extreme case: each example has its own value
    - e.g. example ID; Day attribute in weather data
- Subsets are more likely to be pure if there is a large number of different attribute values
  - Information gain is biased towards choosing attributes with a large number of values



Gain (S, Humidity)
$= .940 - (7/14).985 - (7/14).592$
$= .151$

Gain (S, Wind)
$= .940 - (8/14).811 - (6/14)1.0$
$= .048$

$Gain(S, Temperature) = 0.029$

$Gain(S, Outlook) = 0.246$

- Entropy of split:

$$I(\text{Day}) = \tfrac{1}{14}\left(E([0,1]) + E([0,1]) + \ldots + E([0,1])\right) = 0$$

  - Information gain is maximal for Day (0.940 bits)

Attributes with large # of values

- This may cause several problems:
  - *Overfitting*
    - selection of an attribute that is non-optimal for prediction
  - *Fragmentation*
    - data are fragmented into (too) many small sets

Attributes with large # of values – measure

- Intrinsic information of a split
  - entropy of distribution of $\boxed{\text{instances}}$ into branches
  - i.e. how much information do we need to tell which branch an instance belongs to

$$IntI(S, A) = -\sum_i \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right)$$

- Example:
  - Intrinsic information of Day attribute:

$$IntI(\text{Day}) = 14 \times \left(-\tfrac{1}{14} \cdot \log\left(\tfrac{1}{14}\right)\right) = 3.807$$

- Observation:
  - Attributes with higher intrinsic information are less useful

Gain Ratio

- modification of the information gain that reduces its bias towards multi-valued attributes
- takes number and size of branches into account when choosing an attribute
  - corrects the information gain by taking the *intrinsic information* of a split into account
- Definition of Gain Ratio:

$$GR(S,A)=\frac{Gain(S,A)}{IntI(S,A)}$$

- Example:
  - Gain Ratio of Day attribute

$$GR(\text{Day})=\frac{0.940}{3,807}=0.246$$

Handling numerical attributes

- Standard method: binary splits
  - E.g. Temperature < 78
- Multiple split points possible
- Computationally more demanding

Handling numerical attributes – some optimizations

- Assume a numerical attribute for Temperature
- First step:
  - Sort all examples according to the value of this attribute
  - Could look like this:

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

Handling numerical attributes – some optimizations

- Assume a numerical attribute for Temperature
- First step:
  - Sort all examples according to the value of this attribute
  - Could look like this:

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

- One split between each pair of values
  - E.g. Temperature < 71.5: yes/4, no/2
    Temperature ≥ 71.5: yes/5, no/3

$$I(\text{Temperature} @ 71.5) = \frac{6}{14} \cdot E(\text{Temperature} < 71.5) + \frac{8}{14} E(\text{Temperature} \geq 71.5) = 0.939$$

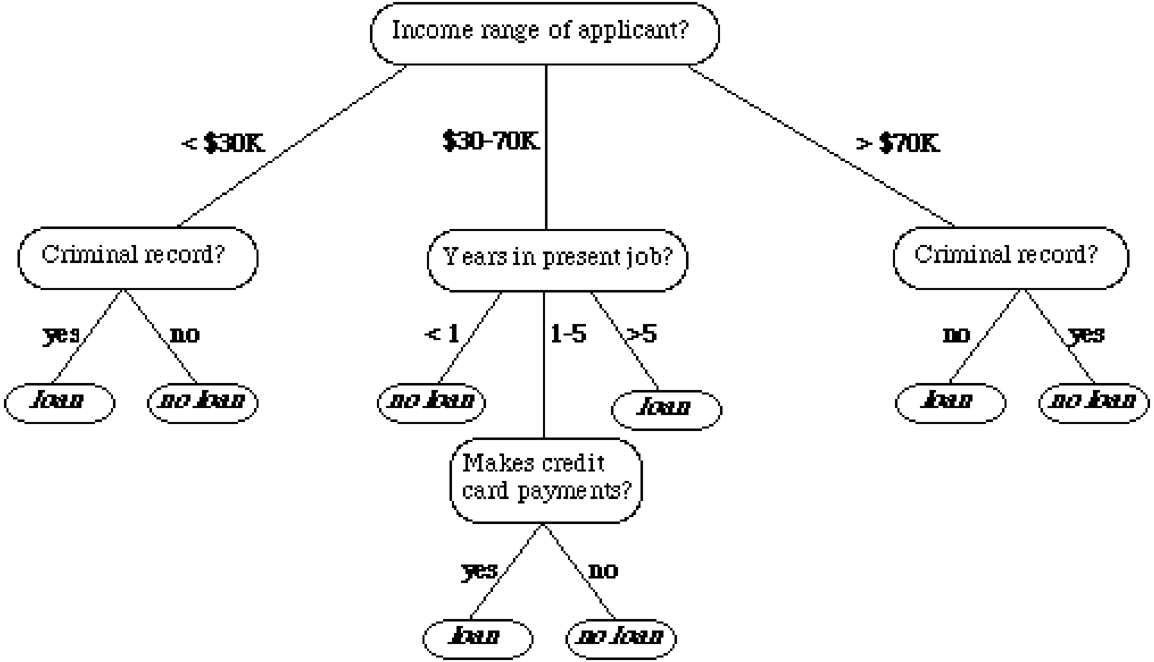
- Split points can be placed between values or directly at values

Handling numerical attributes

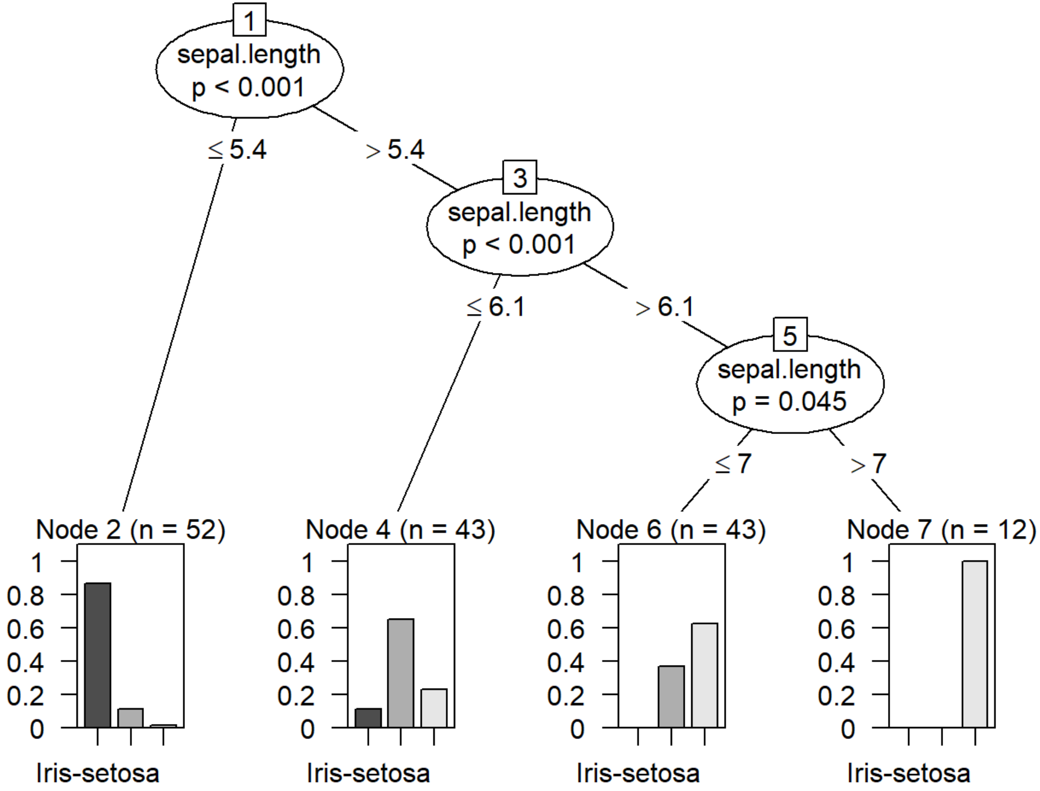
- Splitting (multi-way) on a nominal attribute exhausts all information in that attribute
  - Nominal attribute is tested (at most) once on any path in the tree
- Not so for binary splits on numerical attributes (why ?)
- ➔ Attribute may be tested multiple times in the tree
- → Tree may become hard to read

Handling numerical attributes

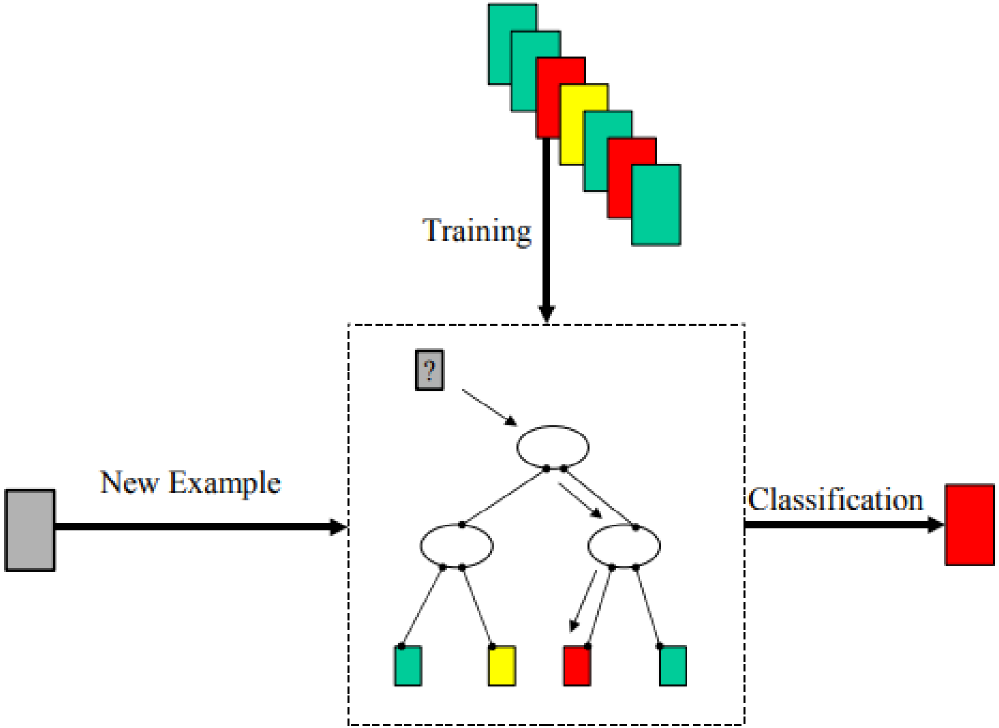
- Discretization / Clustering



Decision Tree with numerical attribute



Deployment



Learning Algorithm for Decision Trees

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

$$\mathbf{x} = (x_1, \ldots, x_d)$$

$$x_j, y \in \{0,1\}$$

GROWTREE(S)

if ($y = 0$ for all $\langle \mathbf{x}, y \rangle \in S$) **return** new leaf(0)

else if ($y = 1$ for all $\langle \mathbf{x}, y \rangle \in S$) **return** new leaf(1)

else

  choose best attribute x_j

  $S_0 =$ all $\langle \mathbf{x}, y \rangle \in S$ with $x_j = 0$;

  $S_1 =$ all $\langle \mathbf{x}, y \rangle \in S$ with $x_j = 1$;

  return new node(x_j, GROWTREE(S_0), GROWTREE(S_1))

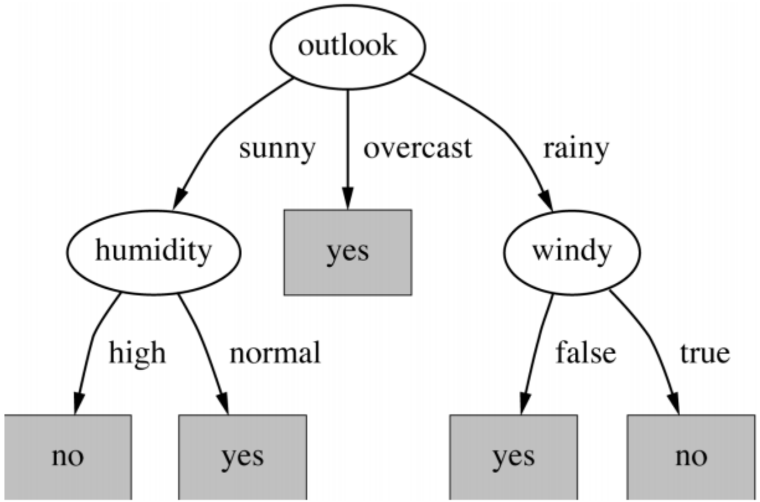
DT algs differ on this choice!
- ID3
- CAT4.5
- CART

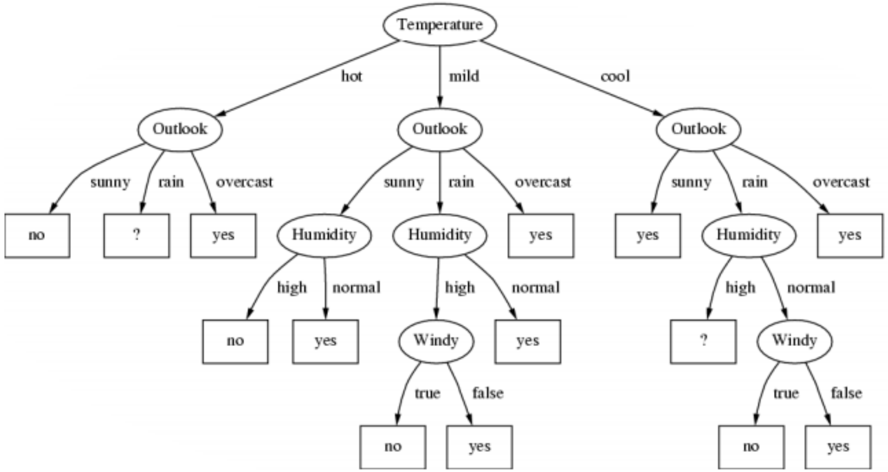
Other issues to address

- Missing attributes
- Attribute values not seen during tree induction (construction)
- Attribute missing in ‘test phase’
  - Divide into pieces etc.

I, Donald J. Trump, am currently inclined to use a position of power to modify the location that is above the country of Mexico and below the country of Canada to be held up to the standards it cemented to be of excellent status, at an earlier point in the country's lifetime.

Small is often better





The Smallest Decision Tree

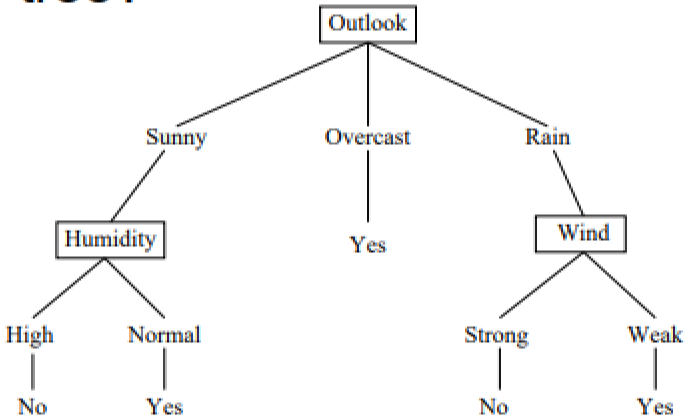
- Learning the smallest DT is NP-hard (Hyafil & Rivest ’76)
- Greedy Heuristic
  - Start from empty decision tree
  - Split on next best attribute (feature)
  - Recurse

Overfitting in Decision Trees

- Consider adding noisy training example #15:

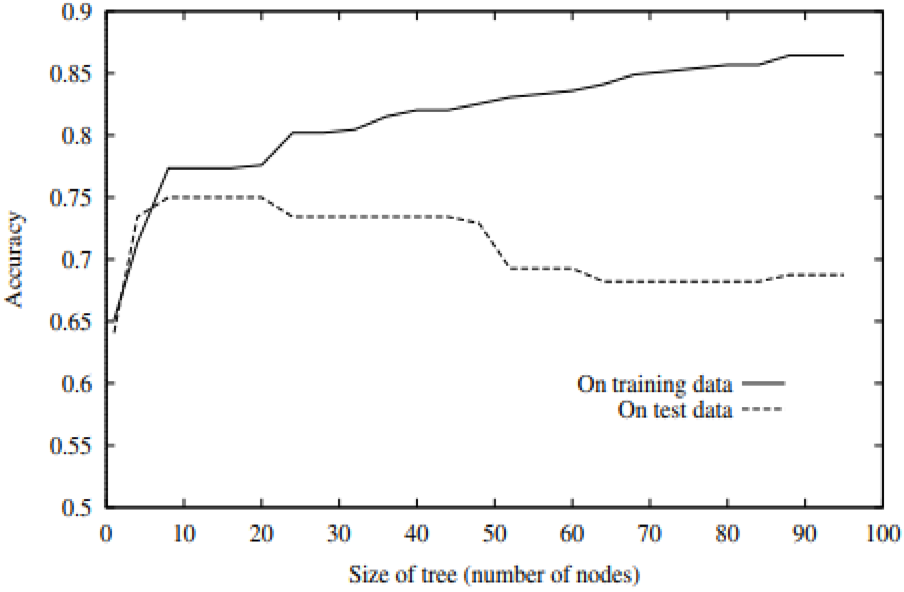
$$Sunny,\ Hot,\ Normal,\ Strong,\ PlayTennis = No$$

- What effect on earlier tree?



Overfitting in Decision Trees

- Overfitting can occur with noisy training examples, and also when small numbers of examples are associated with leaf nodes (→ coincidental or accidental regularities)

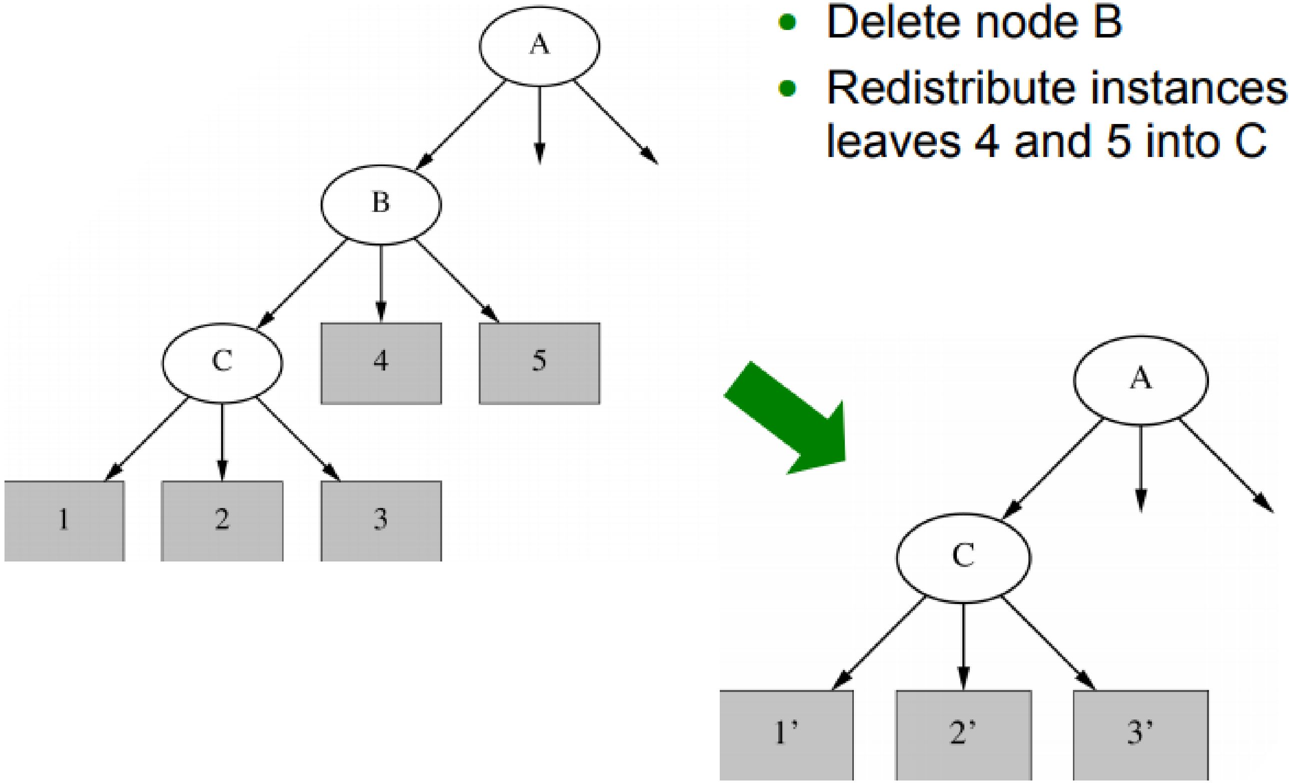


Avoiding overfitting

- Pre-pruning : stop growing tree based on statistical tests of significance
- Post-pruning : Grow full tree, then prune

Post-pruning by subtree raising

- Delete node B
- Redistribute instances of leaves 4 and 5 into C



Decision Trees → Code

rec	Age	Income	Student	Credit_rating	Buys_computer(CLASS)
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31…40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31…40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31…40	Medium	No	Excellent	Yes
r13	31…40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

IF *age* = "<=30" AND *student* = "*no*" THEN
buys_computer = "*no*"

IF *age* = "<=30" AND *student* = "*yes*" THEN
buys_computer = "*yes*"

IF *age* = "31…40" THEN
buys_computer = "*yes*"

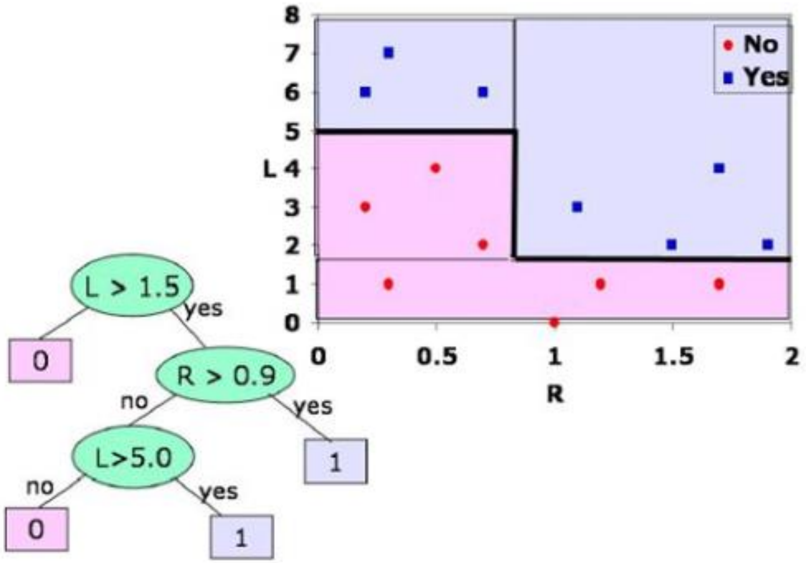
IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN
buys_computer = "no"

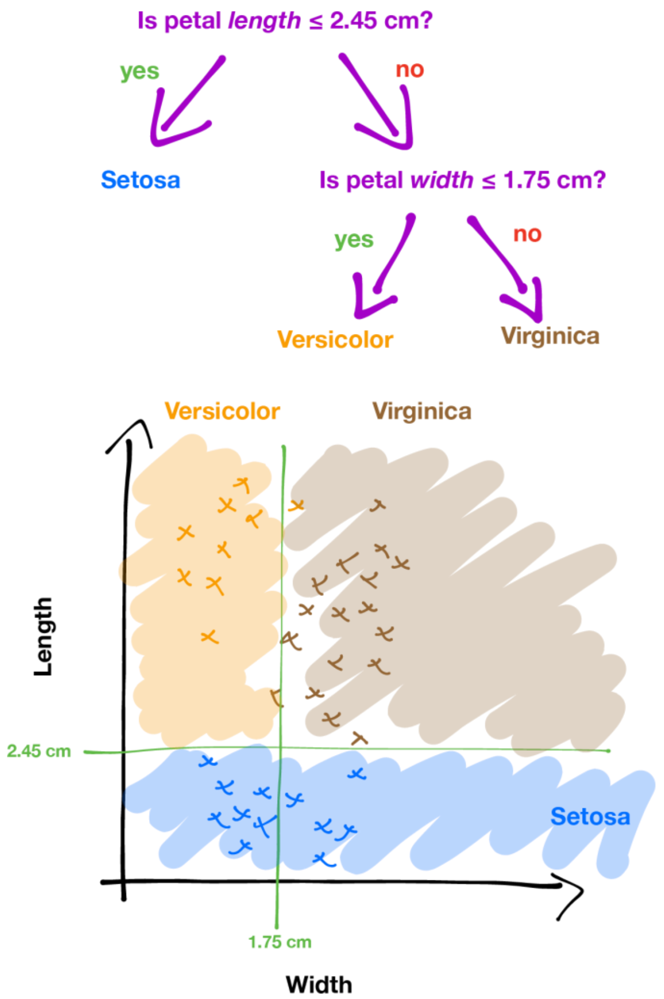
IF *age* = ">40" AND *credit_rating* = "*fair*" THEN
buys_computer = "yes"

Attributes with Differing Costs

- Measuring attribute costs something
  - prefer cheap ones if possible
  - use costly ones only if good gain
- Introduce cost term in selection measure
  - no guarantee in finding optimum, but give bias towards cheapest
- Example applications
  - robot & sonar: time required to position
  - medical diagnosis: cost of a laboratory test

Decision Boundaries





Decision trees for classification

Some real examples (from Russell & Norvig, Mitchell)

- BP’s GasOIL system for separating gas and oil on offshore platforms - decision trees replaced a hand-designed rules system with 2500 rules. C4.5-based system outperformed human experts and saved BP millions. (1986)
- learning to fly a Cessna on a flight simulator by watching human experts fly the simulator (1992)
- can also learn to play tennis, analyze C-section risk, etc.

Advantages of DT

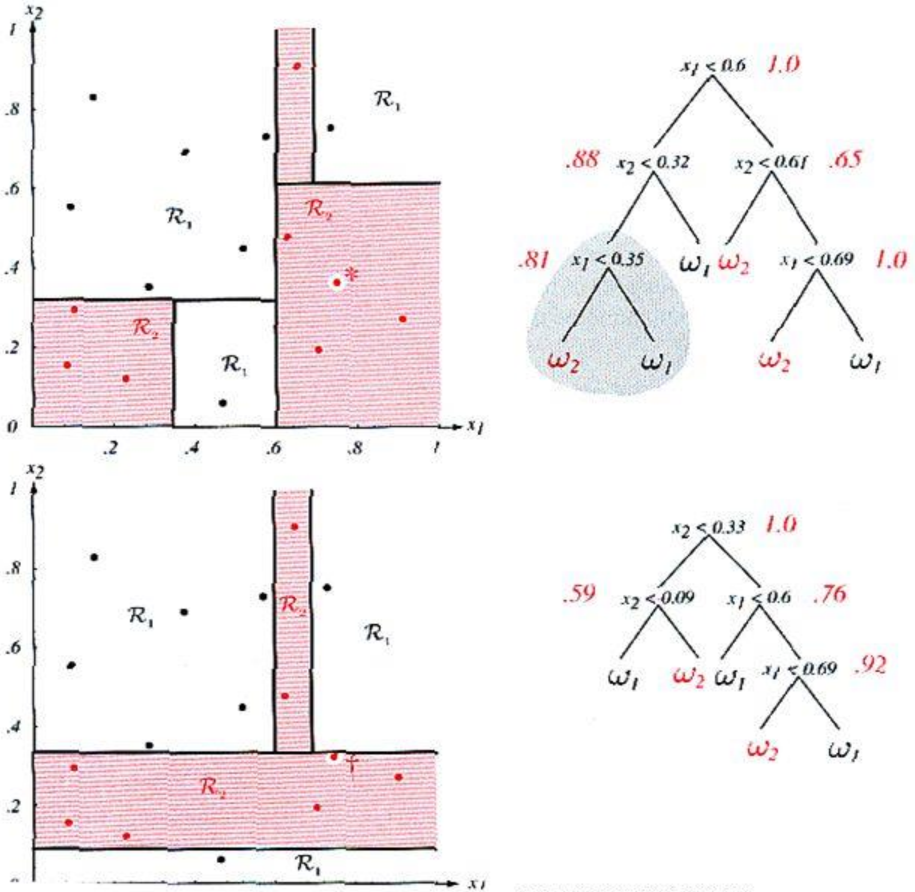
- Easy to use, understand
- Produce rules that are easy to interpret & implement
- Variable selection & reduction is automatic
- Do not require the assumptions of statistical models
- Can work without extensive handling of missing data

Disadvantages

- May not perform well where there is structure in the data that is not well captured by horizontal or vertical splits

- Since the process deals with one variable at a time, no way to capture interactions between variables

Decision Trees are not stable



Moving just one example slightly may lead to quite different trees and space partition!

Lack of stability against small perturbation of data.

Figure from Duda, Hart & Stork, Chap. 8

References and Reading

- https://en.wikipedia.org/wiki/Decision_tree_learning
- Cool demo: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- Entropy in decision trees: https://bricaud.github.io/personal-blog/entropy-in-decision-trees/

- Textbook References
  - [TM] Machine Learning by Tom Mitchell (3.1 – 3.5, 3.7 – 3.8)
  - [PRML] Pattern Recognition and Machine Learning by Chris Bishop (1.2 (intro), 1.6)
  - [DHS] Duda and Hart (8.1 – 8.4)

- Code
  - https://scikit-learn.org/stable/modules/tree.html
  - https://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html