Learning decision trees – full examples

Lecture 2.3

Steps of the tree induction

- **Step 1.** Compute entropy of the instances in the current set (in the beginning – the entire dataset).
- **Step 2.** For each attribute, compute information gain and select the attribute which gives maximum information gain.
- **Step 3.** Create a node with the selected attribute and create branch for each possible attribute value. Split instances into subsets according to this value.
- **Step 4.** For each subset:

  If no split is possible, create leaf node and mark it with the majority class

  Else go to Step 1

Decision tree induction algorithm: *pseudocode*

- ID3 algorithm

current set = all
parent entropy = entropy of *current set*

- **Step 1.**

For each attribute:
    compute entropy
    compute information gain vs. *parent entropy*
best attribute = attribute with maximum information gain

- **Step 2.**

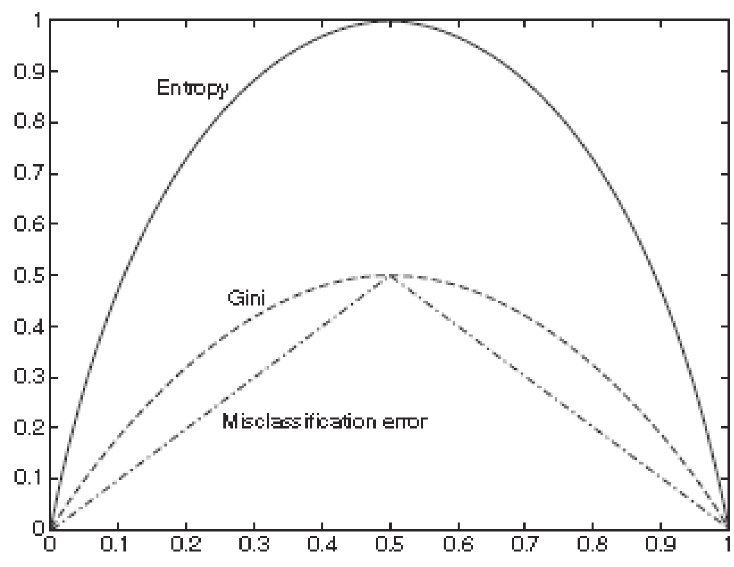
create a node with *best attribute*
create branch for each possible attribute value
split instances into subsets according to the value of *best attribute*

- **Step 3.**

For each *subset*:
    If no split is possible then
        create leaf node
        mark it with the majority class
    Else
        current set =subset
        parent entropy = entropy of *current set*
        go to Step 1

The best attribute to split on

- The GINI score is maximized
  ⇔(1.0-GINI score is minimized)
- The average entropy is minimized
  ⇔(the information gain is maximized)



There are many other attribute selection criteria!
(But almost no difference in accuracy)

- ID3 algorithm
- Design issues
- Split criteria

When to stop splitting

- Not to split: all records are of the same class
- Not to split: all records have the same attribute values
- Not to split: when there is no information gain or information gain is not significant

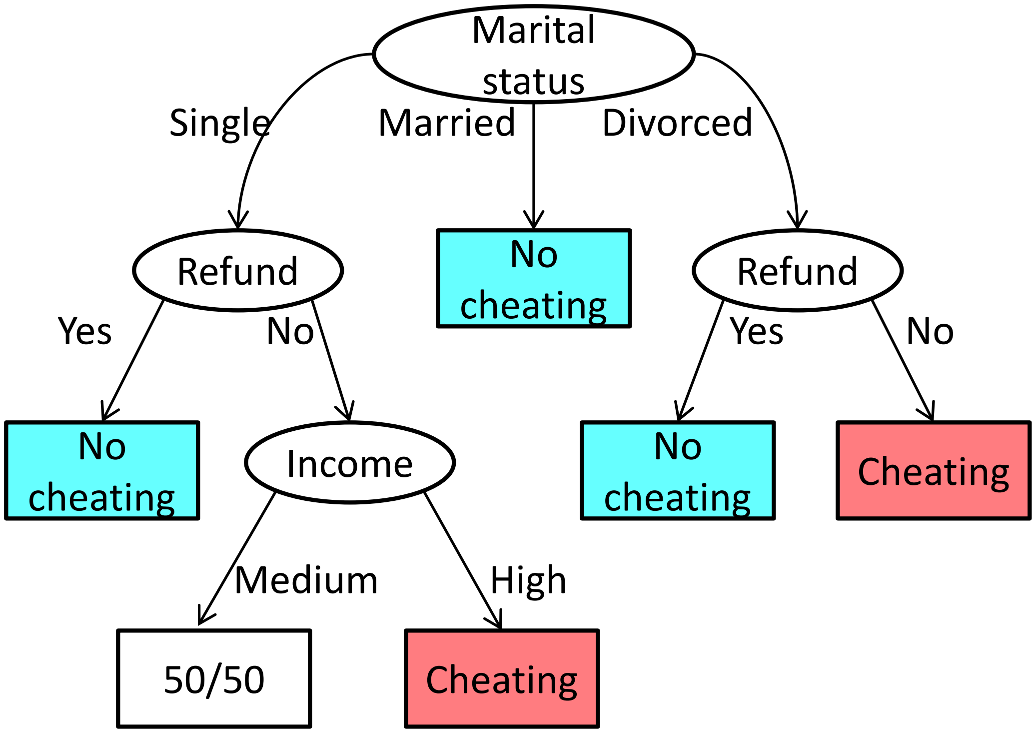
Example 1: Tree induction from tax cheating dataset

ID	Refund	Marital status	Taxable income	Cheat
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

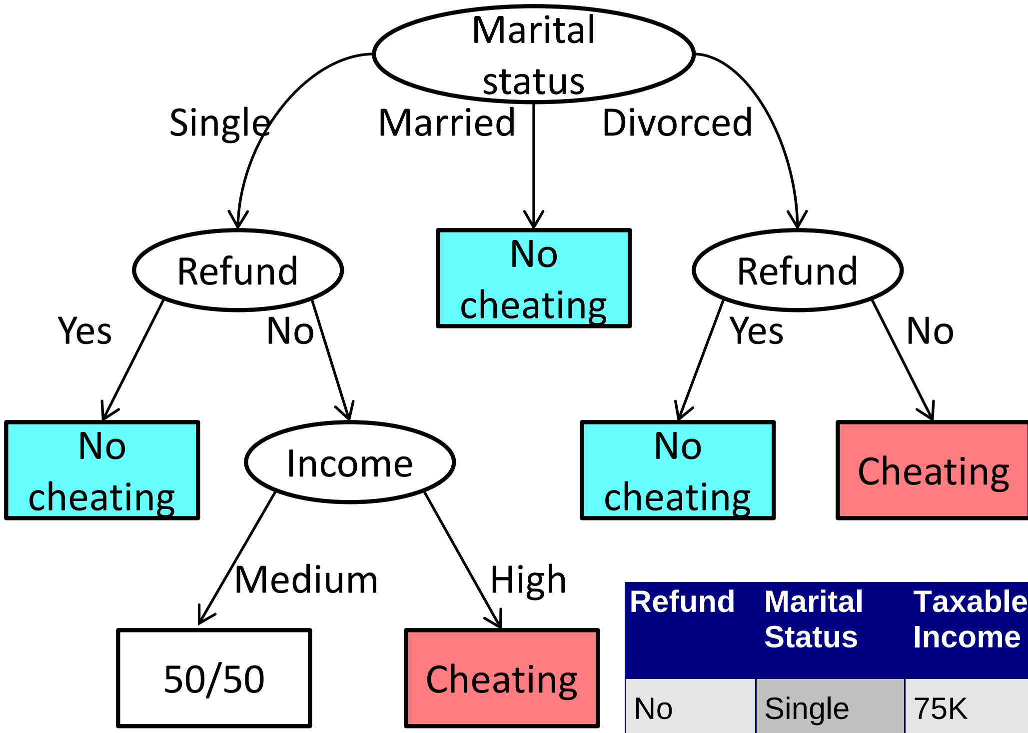
Example 1: Categorizing numeric attributes

ID	Refund	Marital status	Taxable income	Cheat
1	Yes	Single	high	No
2	No	Married	high	No
3	No	Single	medium	No
4	Yes	Married	high	No
5	No	Divorced	medium	Yes
6	No	Married	medium	No
7	Yes	Divorced	high	No
8	No	Single	medium	Yes
9	No	Married	high	No
10	No	Single	high	Yes

Decision tree for tax cheating dataset

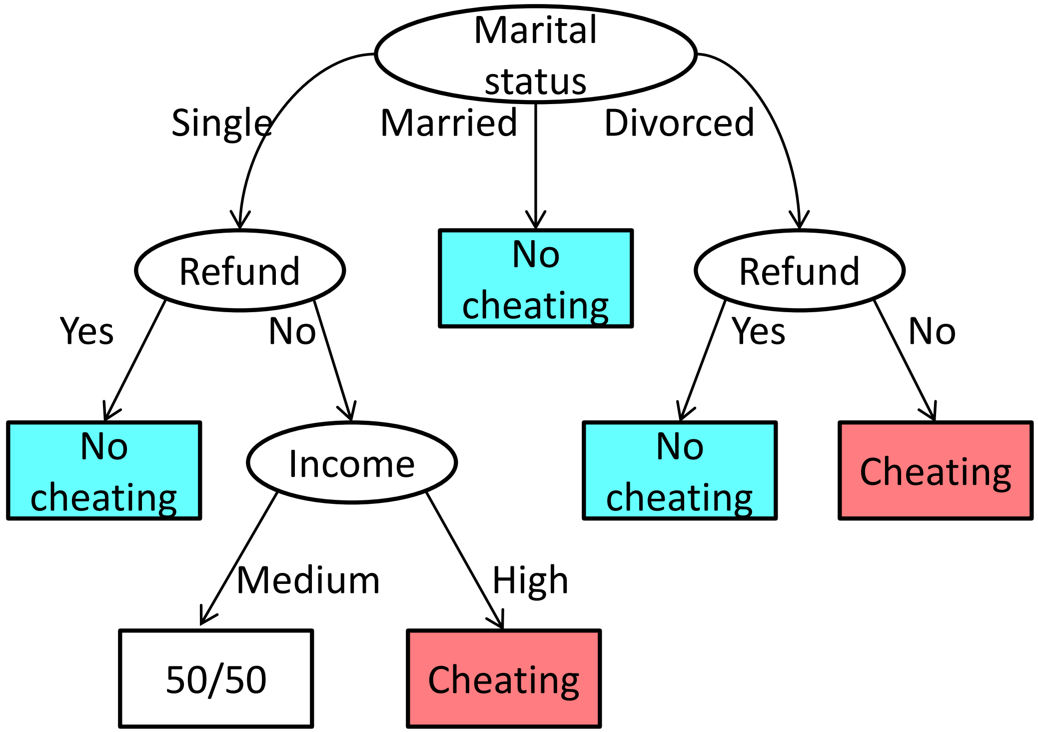


Classify new records



Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Identify the most discriminative attributes

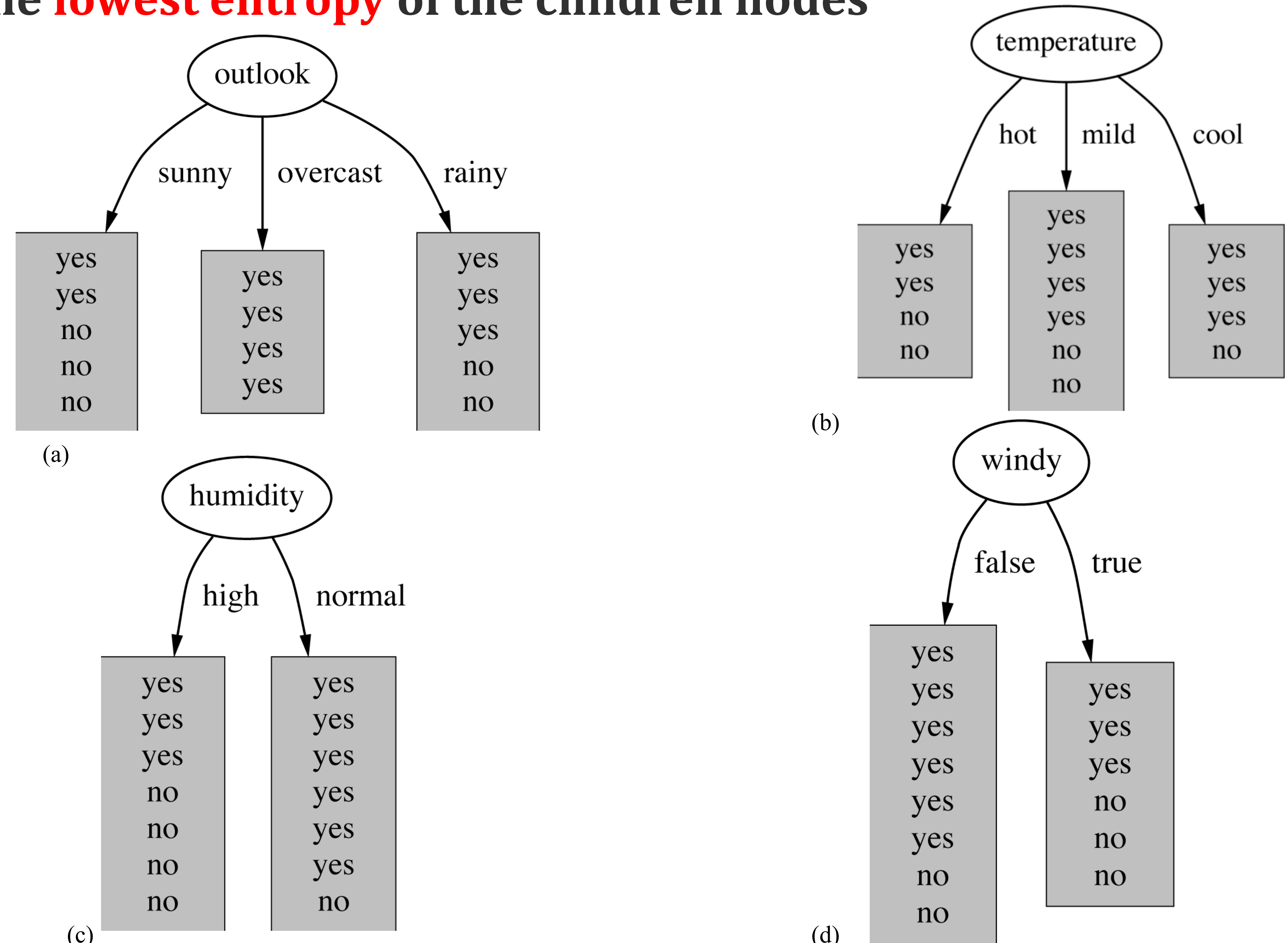


The most important attributes are at the top of the tree

Example 2. Weather data

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

Choose attribute that results in the lowest entropy of the children nodes



Attribute "Outlook"

outlook=sunny

info([2,3]) = entropy(2/5,3/5) = -2/5*log(2/5,2) -3/5*log(3/5,2) = .971

outlook=overcast

info([4,0]) = entropy(4/4,0/4) = -1*log(1,2) -0*log(0,2) = 0

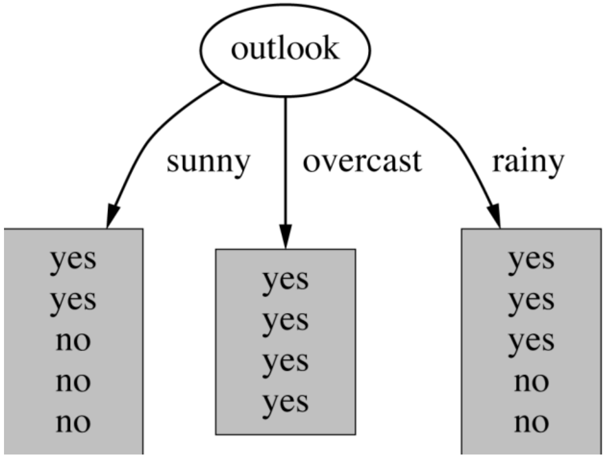
0*log(0) is normally not defined.

outlook=rainy

info([3,2]) = entropy(3/5,2/5) = -3/5*log(3/5,2)-2/5*log(2/5,2) = .971

average entropy:

.971*(5/14) + 0*(4/14) + .971*(5/14) = **.693**



Attribute "Temperature"

temperature=hot

info([2,2]) = entropy(2/4,2/4) = -2/4*log(2/4,2) -2/4*log(2/4,2) = 1

temperature=mild

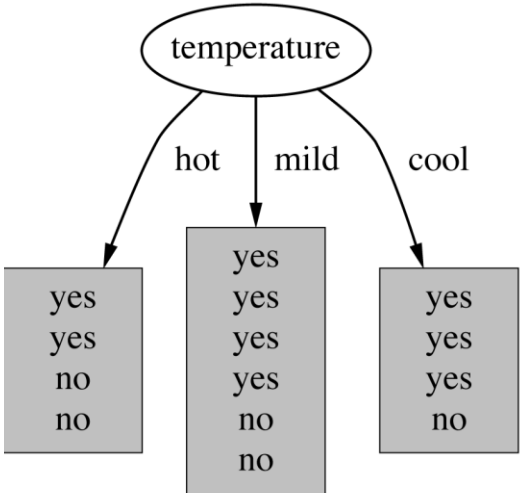
info([4,2]) = entropy(4/6,2/6) = -4/6*log(4/6,2) -2/6*log(2/6,2) = .92

temperature=cool

info([3,1]) = entropy(3/4,1/4) = -3/4*log(3/4,2)-1/4*log(1/4,2) = .811

average entropy:

1*(4/14) + .92*(6/14) + .811*(4/14) = **.91**



Attribute "Humidity"

humidity=high

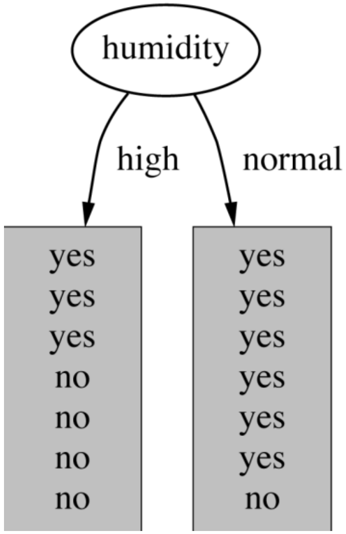
info([3,4]) = entropy(3/7,4/7) = -3/7*log(3/7,2) - 4/7*log(4/7,2) = .985

humidity=normal

info([6,1]) = entropy(6/7,1/7) = -6/7*log(6/7,2) - 1/7*log(1/7,2) = .592

average entropy:

.985*(7/14) + .592*(7/14) = **.788**



Attribute "Windy"

windy=false

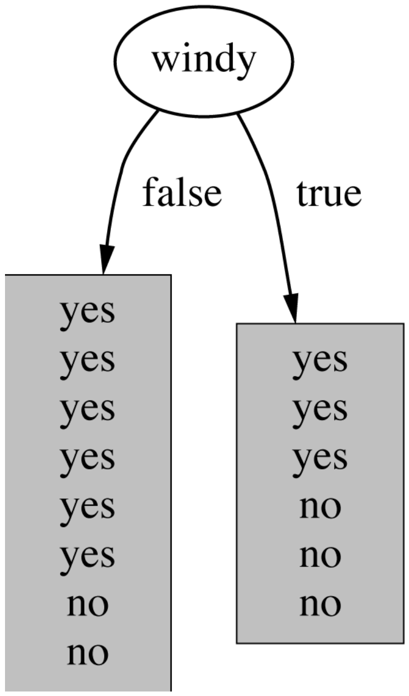
info([6,2]) = entropy(6/8,2/8) = -6/8*log(6/8,2) - 2/8*log(2/8,2) = .811

humidity=true

info([3,3]) = entropy(3/6,3/6) = -3/6*log(3/6,2) - 3/6*log(3/6,2) = 1

average entropy:

.811*(8/14) + 1*(6/14) = **.892**



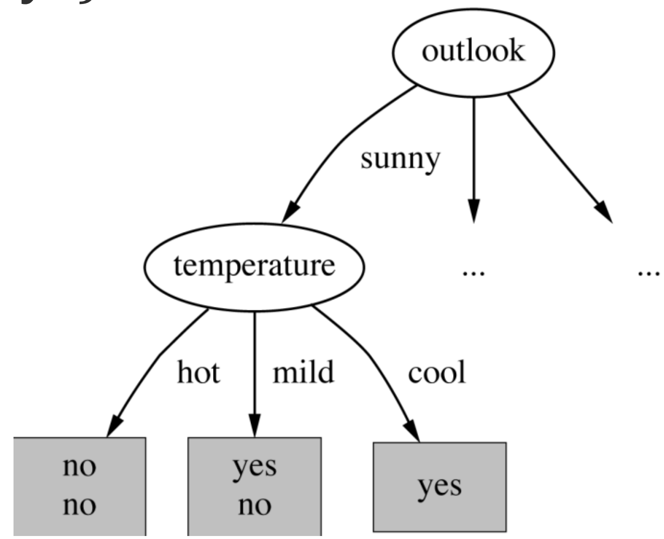
And the winner is...

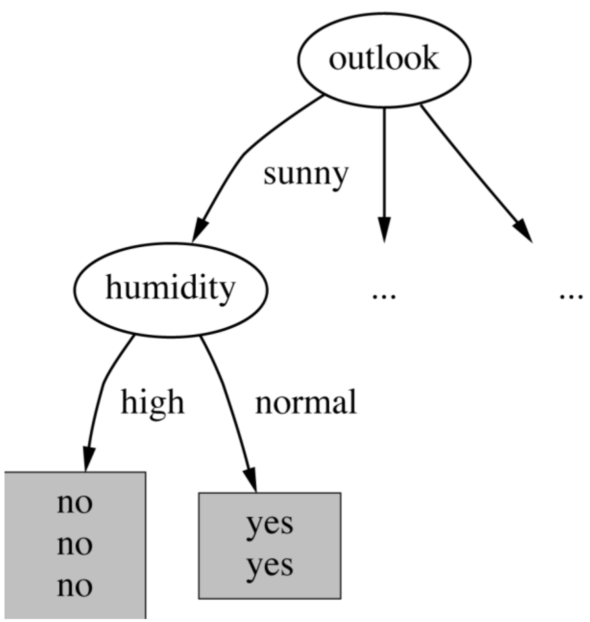
"Outlook"

...So, the root will be "Outlook"

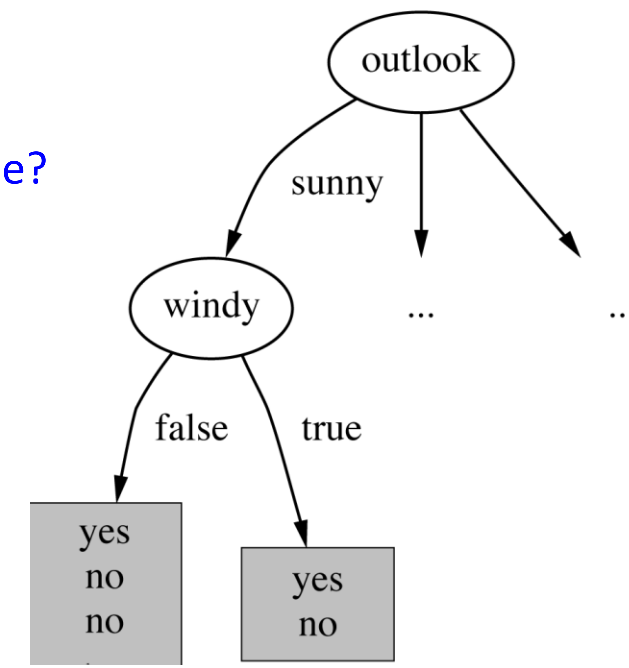
Continuing to split (for Outlook="Sunny")

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

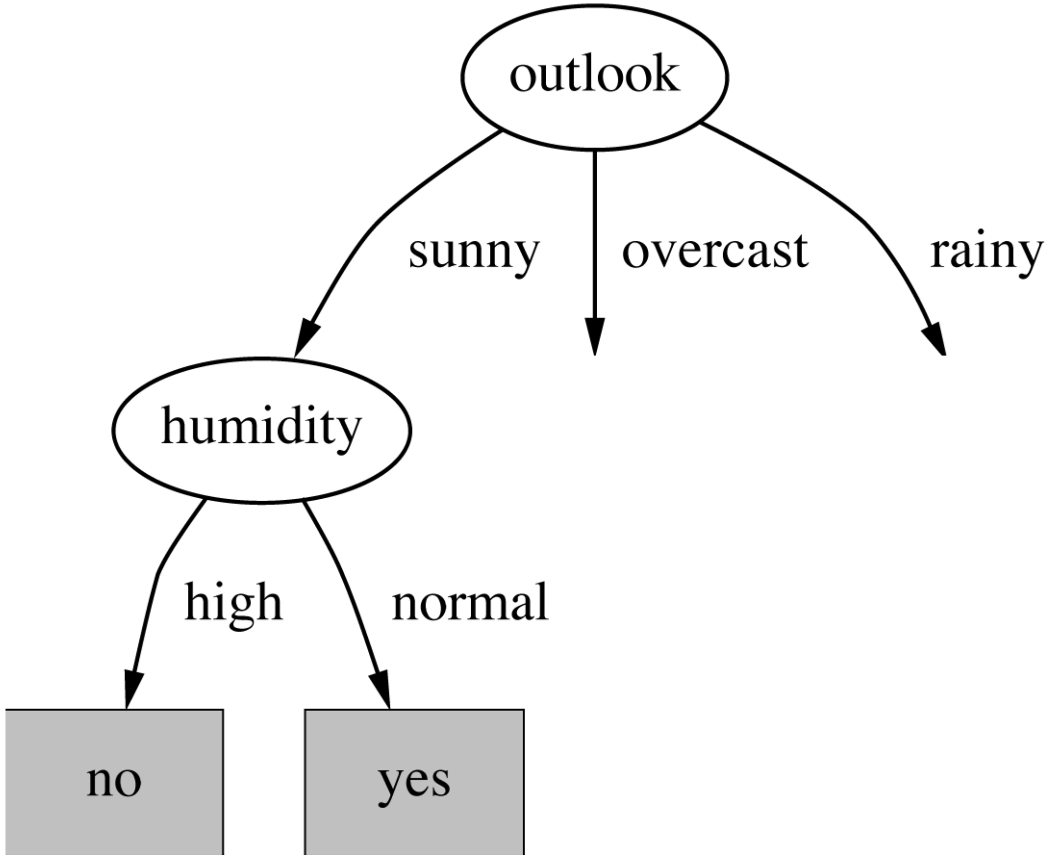




Which one to choose?



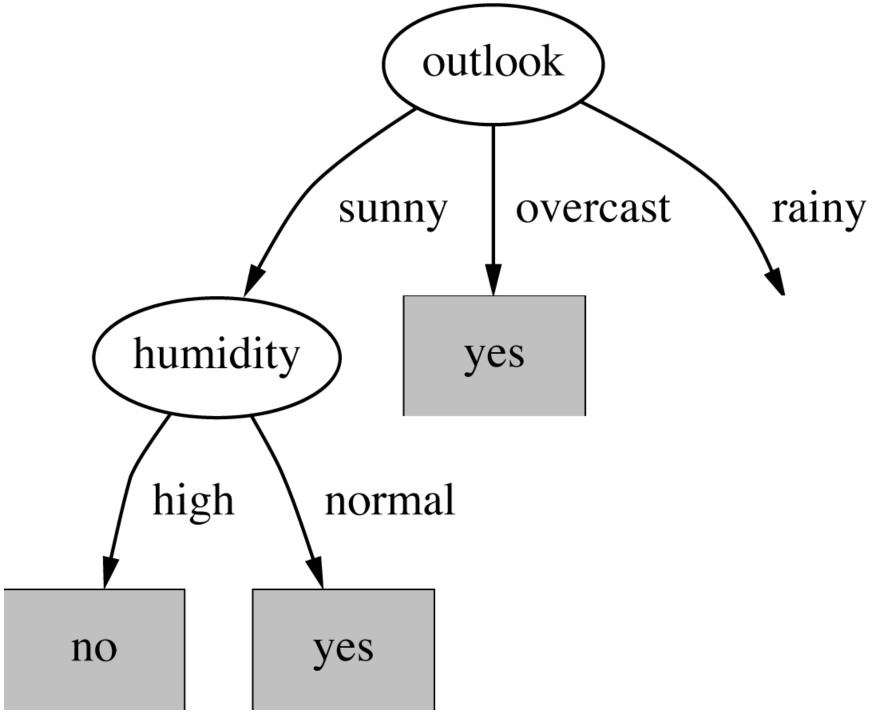
Tree so far



Continuing to split (for Outlook="Overcast")

Outlook	**Temp**	**Humidity**	**Windy**	**Play**
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

- Nothing to split here, "play" is always "yes".



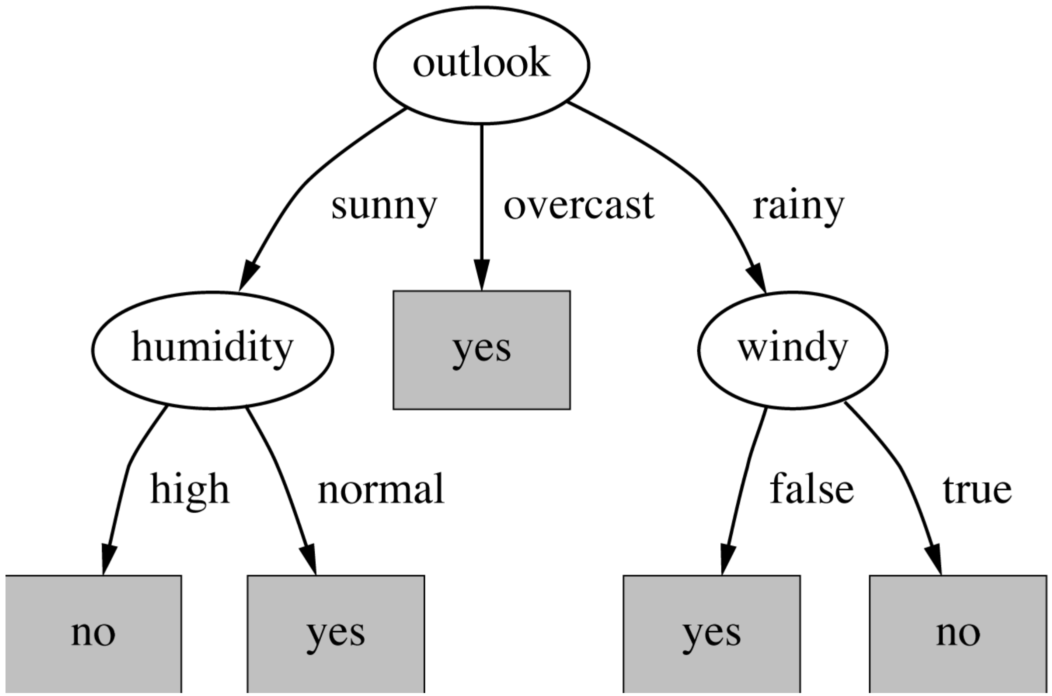
Tree so far

Continuing to split (for Outlook="Rainy")

Outlook	**Temp**	**Humidity**	**Windy**	**Play**
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No

- We see that "Windy" is the one to choose. (**Why?**)

The final decision tree



- Note: not all leaves need to be pure; sometimes identical instances have different classes
- Splitting stops when data can't be split any further or there is no information gain

Example 3: Tree induction from neighbor dataset.

Convert numeric to nominal

Temp	Precip	Day	Clothes	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

Example 3: Tree induction from neighbor dataset

Temp	Precip	Day	Clothes	
warm	None	Fri	Casual	**Walk**
chilly	None	Sun	Casual	**Walk**
chilly	Rain	Wed	Casual	**Walk**
warm	None	Mon	Casual	**Drive**
warm	None	Sat	Formal	**Drive**
warm	None	Sat	Casual	**Drive**
cold	Snow	Mon	Casual	**Drive**
warm	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**