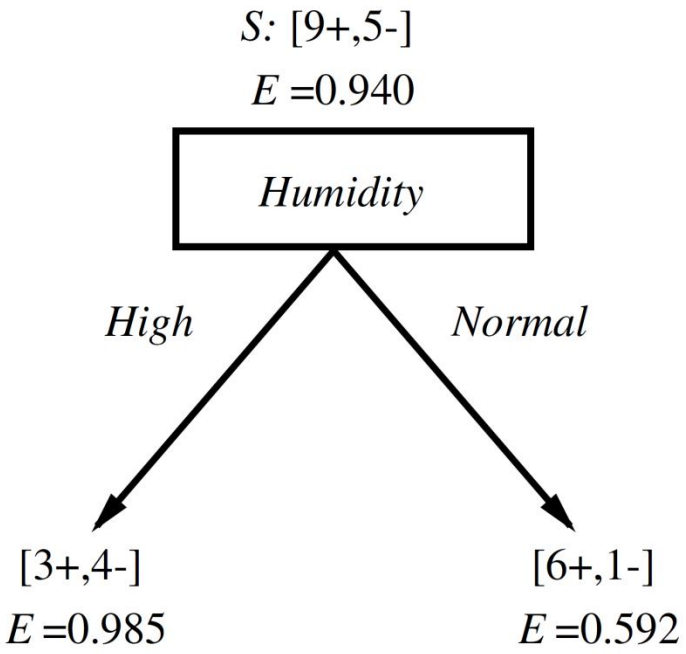
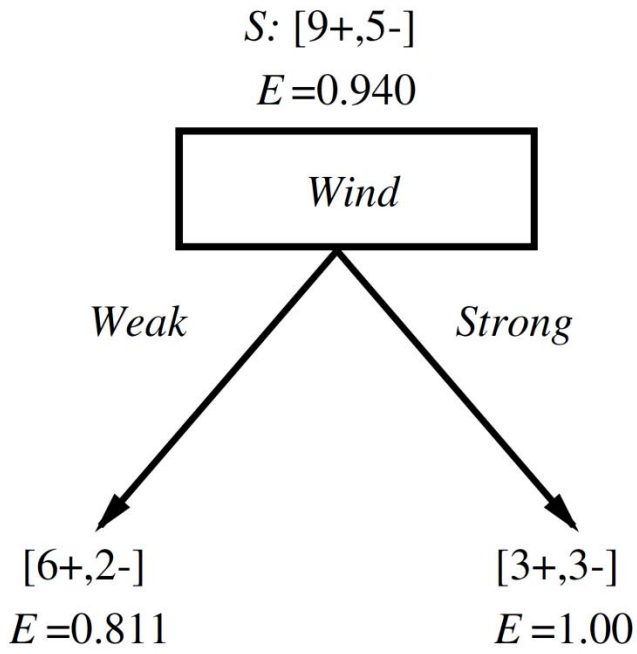
When do I play tennis?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No





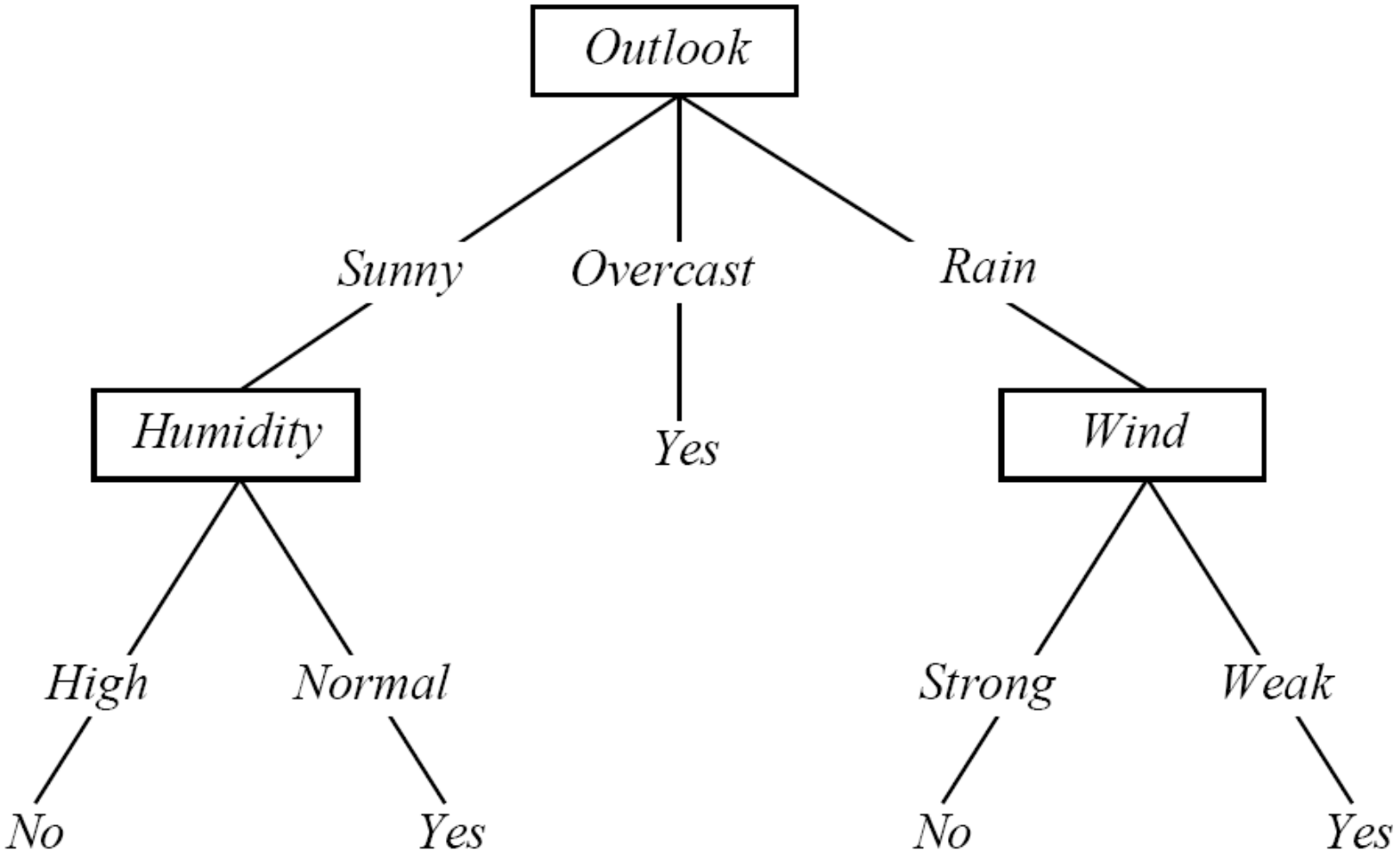
Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

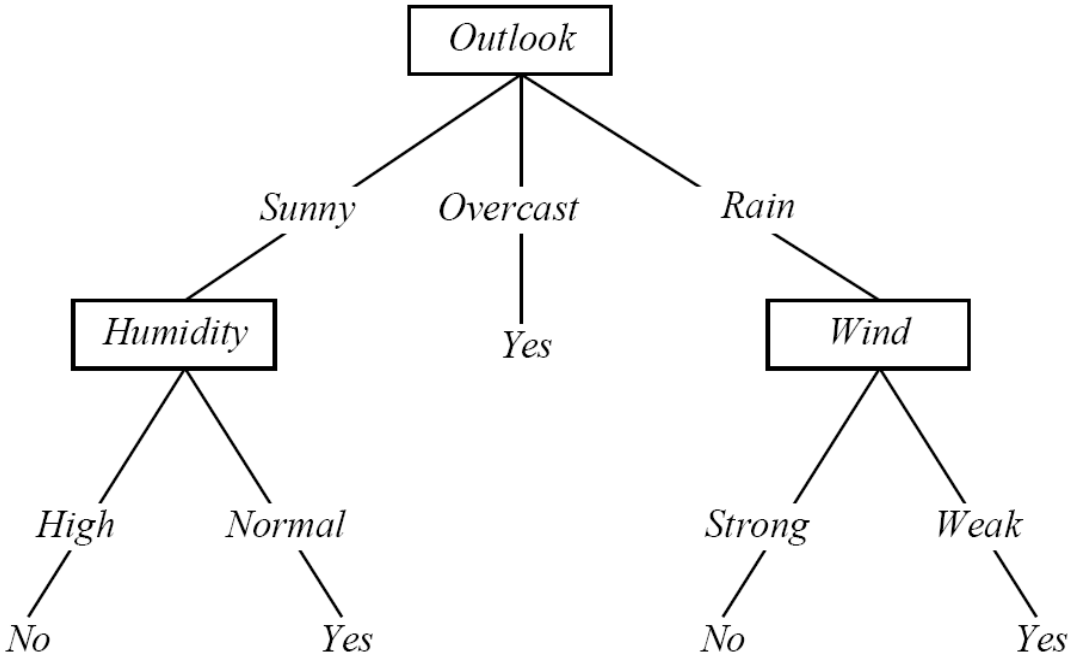
= .940 - (8/14).811 - (6/14)1.0
= .048

Decision Tree



Is the decision tree correct?

- Let's check whether the split on Wind attribute is correct.
- We need to show that Wind attribute has the highest information gain.



When do I play tennis?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Wind attribute – 5 records match

Note: calculate the entropy only on examples that got "routed" in our branch of the tree (Outlook=Rain)

Day					PlayTennis
D1					No
D2					No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Calculation

- $S = \{D4, D5, D6, D10, D14\}$
- Entropy:

  $H(S) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971$
- Information Gain

  $IG(S,Temp) = H(S) - H(S|Temp) = 0.01997$

  $IG(S, Humidity) = H(S) - H(S|Humidity) = 0.01997$

  $IG(S,Wind) = H(S) - H(S|Wind) = 0.971$

Decision Trees: Hypothesis Spaces and Search methods recap

- We search a variable-sized hypothesis space
  - We start at empty and grow it as we build it
- Space is Complete: All target concepts are included in this space
- Local search: No Backtracking
- Batch: At each step, we use all training examples to make a statistically based decision.

Attributes with Many Values

Problem:

- If attribute has many values, $Gain$ will select it
- Imagine using $Date = Jun\_3\_1996$ as attribute

One approach: use $GainRatio$ instead

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

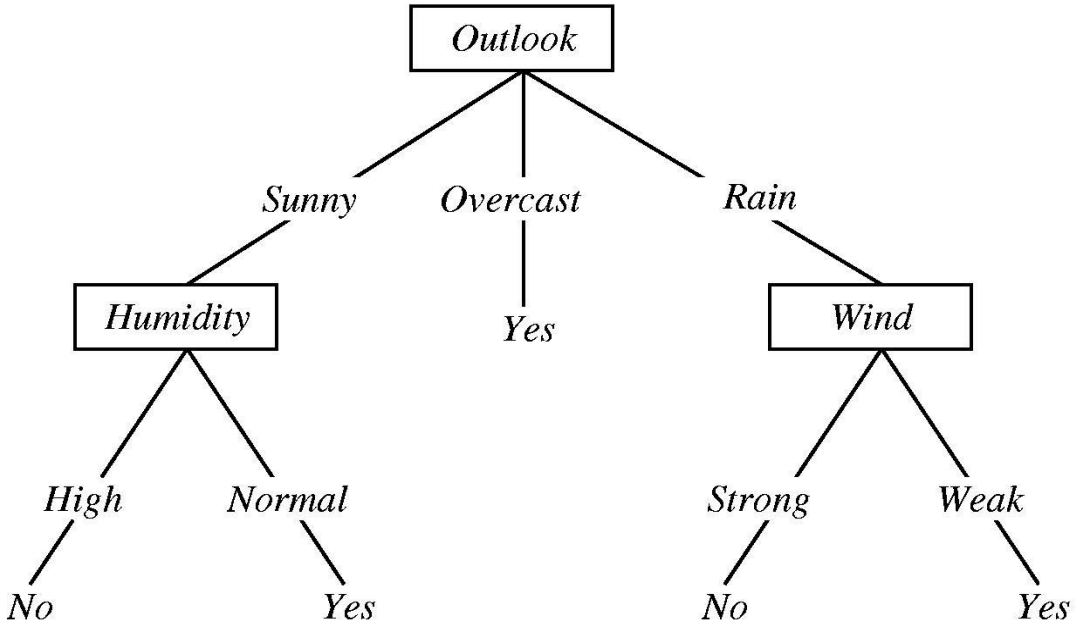
$$SplitInformation(S, A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

where S_i is subset of S for which A has value v_i

A deeper look at Gain Ratio

- What is Split In Information of Date?
  - Data set size n, each has a different date.
- What is Split In Information of an attribute "Weather=Snowing" in Texas?
  - Snows one day and is sunny or overcast on others
- Heuristic
  - First compute Gain
  - Apply Gain ratio only on attributes which have above average Gain.

Overfitting in Decision Trees

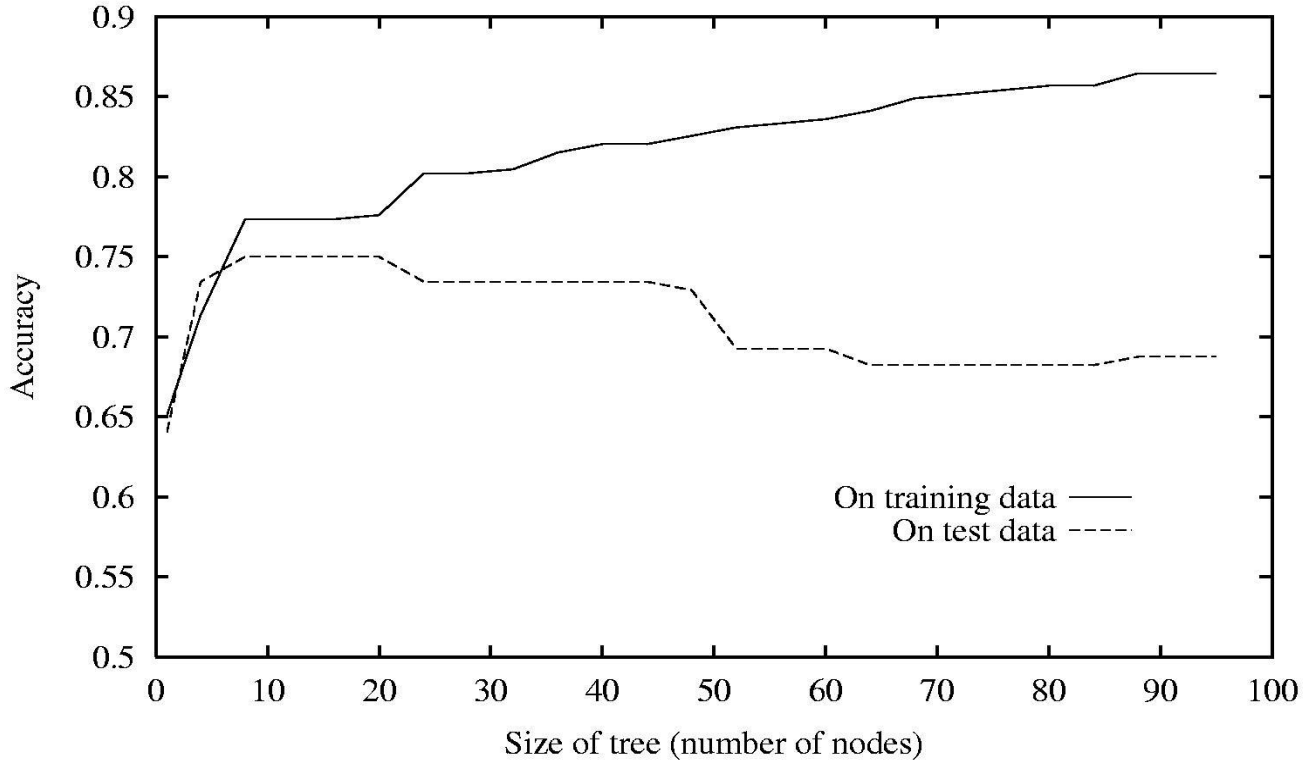


Consider adding a noisy training example:

Sunny, Hot, Normal, Strong, PlayTennis=No

What effect on tree?

Overfitting in Decision Tree Learning



Sources of Overfitting

- Noise
- Small number of examples associated with each leaf
  - What if only one example is associated with a leaf. Can you believe it?
  - Coincidental regularities
- Generalization is the most important criteria
  - Your method should work well on examples which you have not seen before.

Avoiding Overfitting

- Two approaches
  - Stop growing the tree when data split is not statistically significant
  - Grow tree fully, then post-prune
- Key Issue: What is the correct tree size?
  - Divide data into training and validation set
    - Random noise in two sets might be different
  - Apply statistical test to estimate whether expanding a particular node is likely to produce an improvement beyond the training set
  - Add a complexity penalty

Reduced-Error Pruning

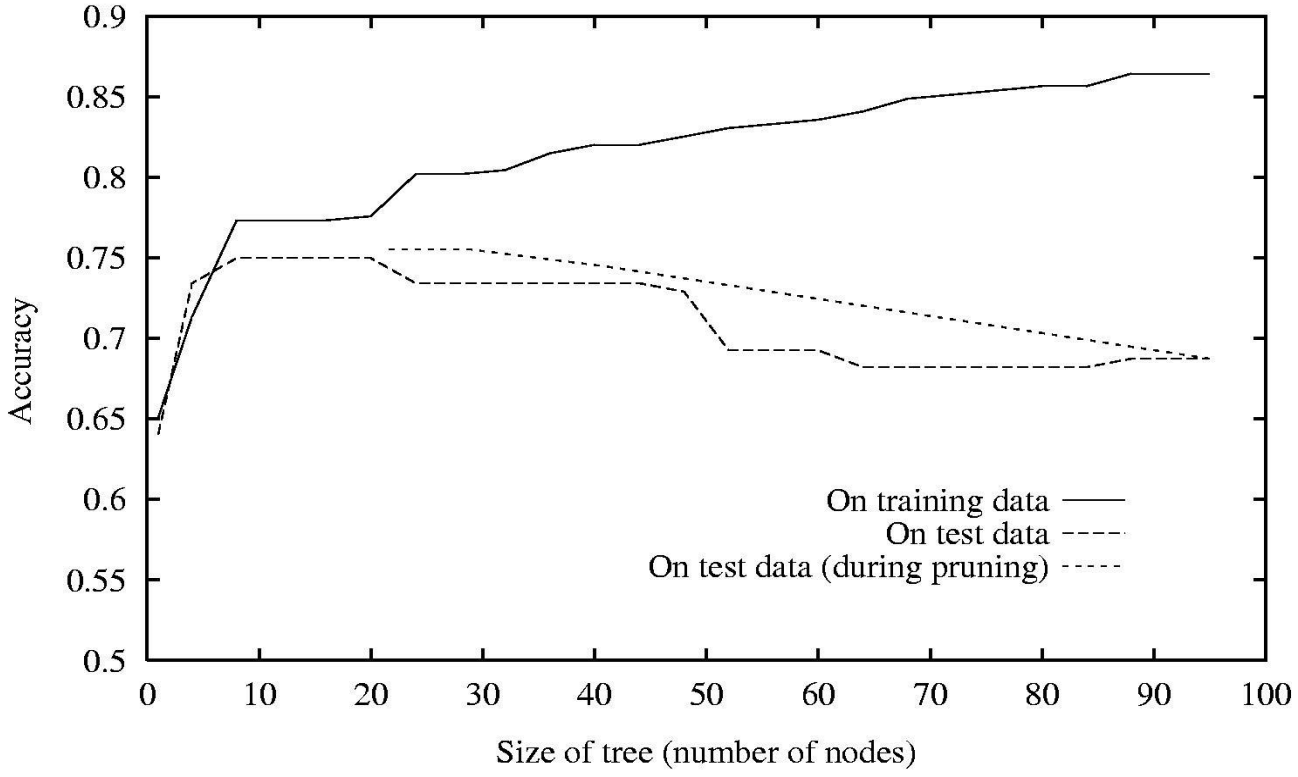
Split data into *training* and *validation* set

Do until further pruning is harmful:

1. Evaluate impact on *validation* set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves *validation* set accuracy

Leaf nodes added because of coincidental regularities are likely to be pruned because the same coincidences are unlikely to occur in the validation set

Effect of Reduced-Error Pruning

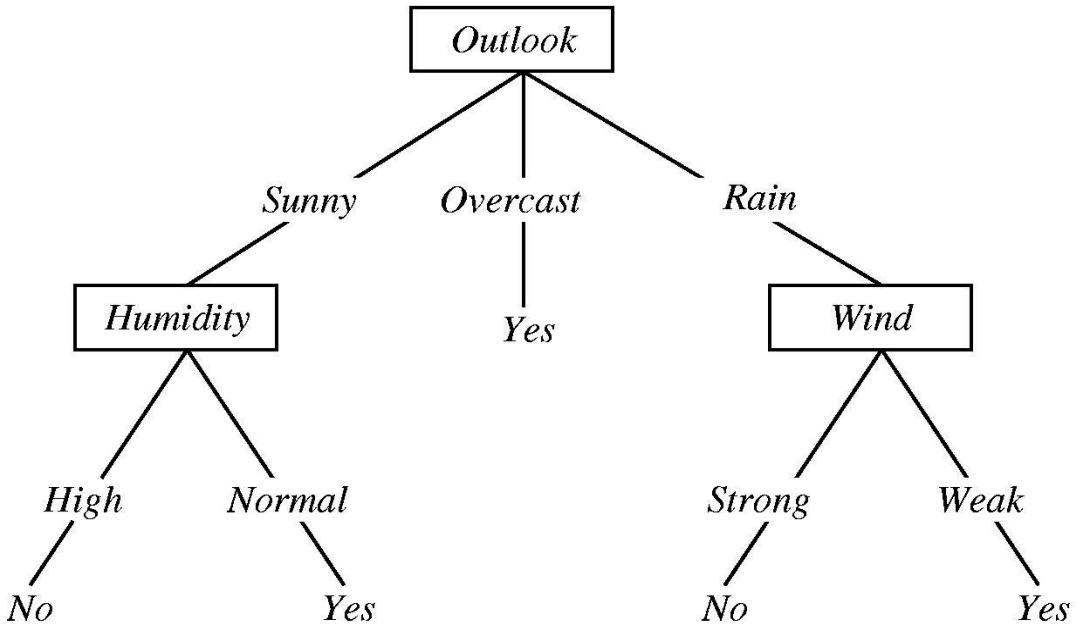


Rule Post-Pruning

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

Perhaps most frequently used method (e.g., C4.5)

Converting A Tree to Rules



IF $(Outlook = Sunny)\ AND\ (Humidity = High)$
THEN $PlayTennis = No$

IF $(Outlook = Sunny)\ AND\ (Humidity = Normal)$
THEN $PlayTennis = Yes$

...

Handling Missing Values

- Missing values: Some attribute-values in an example are missing
  - Example: patient data. You don't expect blood test results for everyone.
- Treat the missing value as another value
- Ignore instances having missing values
  - Problematic because you are throwing away important data. Data is scarce.

Handling Missing Values

- Probabilistic approach
  - Assign a probability to each possible value of A
  - Let us assume that A(x=1)=0.4 and A(x=0)=0.6
  - A fractional 0.4 of instance goes to branch A(x=1) and 0.6 to branch A(x=0)
  - Use fractional instances to compute gain
- Classification
  - Most probable classification

Handling Continuous attributes

- Thresholding
- How to select Thresholds?

40	48	60	72	80	90
no	no	yes	yes	yes	no

- Pick a threshold that has the highest gain!
- Sort the examples and calculate gain at all points where the classification changes from "yes to no" or "no to yes"
  - Provably maximizes the information gain.

Summary: Decision Trees

- Representation
- Tree growth
- Choosing the best attribute
- Overfitting and pruning
- Special cases: Missing Attributes and Continuous Attributes
- Many forms in practice: CART, ID3, C 4.5