

# CS 188: Artificial Intelligence

## Spring 2006

### Lecture 11: Decision Trees

2/21/2006

Dan Klein – UC Berkeley

Many slides from either Stuart Russell or Andrew Moore

## Today

---

- Formalizing Learning
  - Consistency
  - Simplicity
- Decision Trees
  - Expressiveness
  - Information Gain
  - Overfitting

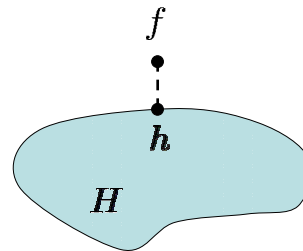
# Inductive Learning (Science)

- Simplest form: learn a function from examples

- A target function:  $f$
- Examples: input-output pairs  $(x, f(x))$
- E.g.  $x$  is an email and  $f(x)$  is spam / ham
- E.g.  $x$  is a house and  $f(x)$  is its selling price

- Problem:

- Given a hypothesis space  $H$
- Given a training set of examples  $x_i$
- Find a hypothesis  $h(x)$  such that  $h \sim f$



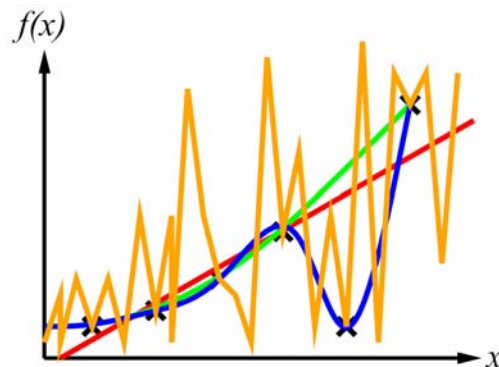
- Includes:

- Classification (multinomial outputs)
- Regression (real outputs)

- How do perceptron and naïve Bayes fit in? ( $H, f, h$ , etc.)

# Inductive Learning

- Curve fitting (regression, function approximation):



- Consistency vs. simplicity
- Ockham's razor

# Consistency vs. Simplicity

- Fundamental tradeoff: bias vs. variance, etc.
- Usually algorithms prefer consistency by default (why?)
- Several ways to operationalize “simplicity”
  - Reduce the hypothesis space
    - Assume more: e.g. independence assumptions, as in naïve Bayes
    - Have fewer, better features / attributes: feature selection
    - Other structural limitations (decision lists vs trees)
  - Regularization
    - Smoothing: cautious use of small counts
    - Many other generalization parameters (pruning cutoffs today)
    - Hypothesis space stays big, but harder to get to the outskirts

# Reminder: Features

- Features, aka attributes
  - Sometimes: TYPE=French
  - Sometimes:  $f_{\text{TYPE=French}}(x) = 1$

Example	Attributes		Target
		<i>Est</i>	<i>WillWait</i>
$X_1$		0-10	T
$X_2$		30-60	F
$X_3$		0-10	T
$X_4$		10-30	T
$X_5$		>60	F
$X_6$		0-10	T
$X_7$		0-10	F
$X_8$		0-10	T
$X_9$		>60	F
$X_{10}$		10-30	F
$X_{11}$		0-10	F
$X_{12}$		30-60	T

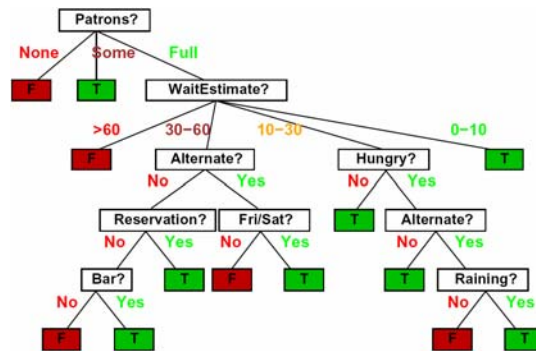
# Decision Trees

- Compact representation of a function:

- Truth table
- Conditional probability table
- Regression values

- True function

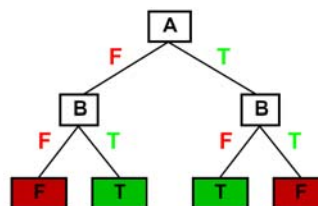
- Realizable: in  $H$



## Expressiveness of DTs

- Can express any function of the features

A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



$$P(C|A, B)$$

- However, we hope for compact trees

# Comparison: Perceptrons

- What is the expressiveness of a perceptron over these features?

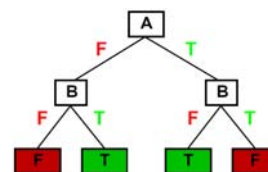
Example	Attributes										Target
	<i>Alt</i>	<i>Bar</i>	<i>Fri</i>	<i>Hun</i>	<i>Pat</i>	<i>Price</i>	<i>Rain</i>	<i>Res</i>	<i>Type</i>	<i>Est</i>	<i>WillWait</i>
$X_1$	<i>T</i>	<i>F</i>	<i>F</i>	<i>T</i>	<i>Some</i>	<i>\$\$\$</i>	<i>F</i>	<i>T</i>	<i>French</i>	<i>0-10</i>	<i>T</i>
$X_2$	<i>T</i>	<i>F</i>	<i>F</i>	<i>T</i>	<i>Full</i>	<i>\$</i>	<i>F</i>	<i>F</i>	<i>Thai</i>	<i>30-60</i>	<i>F</i>

- DTs automatically conjoin features / attributes
  - Features can have different effects in different branches of the tree!
- For a perceptron, a feature's contribution is either positive or negative
  - If you want one feature's effect to depend on another, you have to add a new conjunction feature
  - E.g. adding "PATRONS=full  $\wedge$  WAIT = 60" allows a perceptron to model the interaction between the two atomic features
- Difference between modeling relative evidence weighting (NB) and complex evidence interaction (DTs)
  - Though if the interactions are too complex, may not find the DT greedily

# Hypothesis Spaces

- How many distinct decision trees with  $n$  Boolean attributes?

= number of Boolean functions over  $n$  attributes  
 = number of distinct truth tables with  $2^n$  rows  
 =  $2^{(2^n)}$   
 E.g., with 6 Boolean attributes, there are  
 18,446,744,073,709,551,616 trees



- How many trees of depth 1 (decision stumps)?

= number of Boolean functions over 1 attribute  
 = number of truth tables with 2 rows, times  $n$   
 =  $4n$   
 E.g. with 6 Boolean attributes, there are 24 decision stumps

- More expressive hypothesis space:

- Increases chance that target function can be expressed (good)
- Increases number of hypotheses consistent with training set (bad, why?)
- Means we can get better predictions (lower bias)
- But we may get worse predictions (higher variance)

# Decision Tree Learning

- Aim: find a small tree consistent with the training examples
- Idea: (recursively) choose “most significant” attribute as root of (sub)tree

```
function DTL(examples, attributes, default) returns a decision tree
  if examples is empty then return default
  else if all examples have the same classification then return the classification
  else if attributes is empty then return MODE(examples)
  else
    best ← CHOOSE-ATTRIBUTE(attributes, examples)
    tree ← a new decision tree with root test best
    for each value vi of best do
      examplesi ← {elements of examples with best = vi}
      subtree ← DTL(examplesi, attributes − best, MODE(examples))
      add a branch to tree with label vi and subtree subtree
    return tree
```

## Choosing an Attribute

- Idea: a good attribute splits the examples into subsets that are (ideally) “all positive” or “all negative”



- So: we need a measure of how “good” a split is, even if the results aren’t perfectly separated out

# Entropy and Information

- **Information** answers questions
  - The more uncertain about the answer initially, the more information in the answer
  - Scale: bits
    - Answer to Boolean question with prior  $\langle 1/2, 1/2 \rangle$ ?
    - Answer to 4-way question with prior  $\langle 1/4, 1/4, 1/4, 1/4 \rangle$ ?
    - Answer to 4-way question with prior  $\langle 0, 0, 0, 1 \rangle$ ?
    - Answer to 3-way question with prior  $\langle 1/2, 1/4, 1/4 \rangle$ ?
- A probability  $p$  is typical of:
  - A uniform distribution of size  $1/p$
  - A code of length  $\log 1/p$

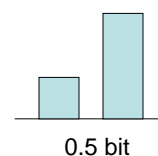
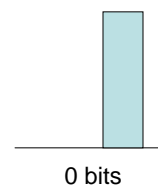
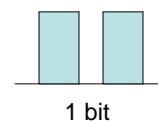
## Entropy

- General answer: if prior is  $\langle p_1, \dots, p_n \rangle$ :
  - Information is the expected code length

$$H(\langle p_1, \dots, p_n \rangle) = E_p \log_2 1/p_i$$

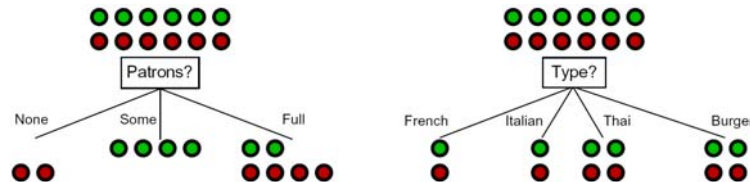
$$= \sum_{i=1}^n -p_i \log_2 p_i$$

- Also called the **entropy** of the distribution
  - More uniform = higher entropy
  - More values = higher entropy
  - More peaked = lower entropy
  - Rare values almost “don’t count”



# Information Gain

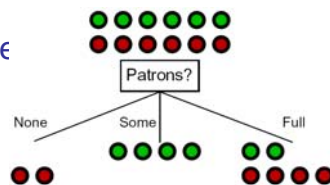
- Back to decision trees!
- For each split, compare entropy before and after
  - Difference is the **information gain**
  - Problem: there's more than one distribution after split!



- Solution: use **expected entropy**, weighted by the number of examples
- Note: hidden problem here! Gain needs to be adjusted for large-domain splits – why?

## Next Step: Recurse

- Now we need to keep growing the
- Two branches are done (why?)
- What to do under “full”?
  - See what examples are there...

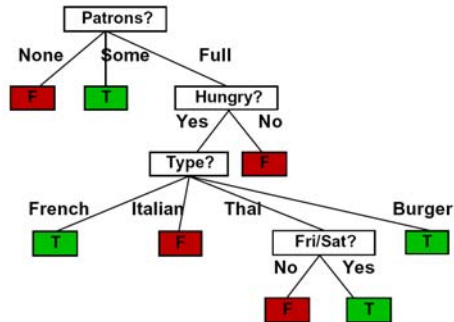


Example	Attributes										Target	
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	Will	Wait
X <sub>1</sub>	T	F	F	T	Some	\$\$\$	F	T	French	0-10	T	
X <sub>2</sub>	T	F	F	T	Full	\$	F	F	Thai	30-40	F	
X <sub>3</sub>	F	T	F	F	Some	\$	F	F	Burger	0-10	T	
X <sub>4</sub>	T	F	T	T	Full	\$	F	F	Thai	10-30	T	
X <sub>5</sub>	T	F	F	F	Full	\$\$\$	F	T	French	60-80	F	
X <sub>6</sub>	F	T	F	T	Some	\$\$	T	T	Italian	0-10	T	
X <sub>7</sub>	F	T	F	F	None	\$	T	F	Burger	0-10	F	
X <sub>8</sub>	F	F	F	T	Some	\$\$	T	T	Thai	0-10	T	
X <sub>9</sub>	F	T	T	F	Full	\$	F	F	Burger	60-80	F	
X <sub>10</sub>	T	T	T	T	Full	\$\$\$	F	T	Italian	10-30	F	
X <sub>11</sub>	F	F	F	F	None	\$	F	F	Thai	0-10	F	
X <sub>12</sub>	F	T	T	T	Full	\$	F	F	Burger	10-30	T	



## Example: Learned Tree

- Decision tree learned from these 12 examples:



- Substantially simpler than “true” tree
  - A more complex hypothesis isn't justified by data
- Also: it's reasonable, but wrong

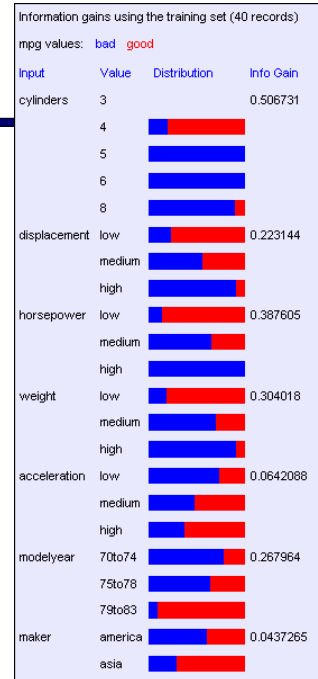
## Example: Miles Per Gallon

40 Examples

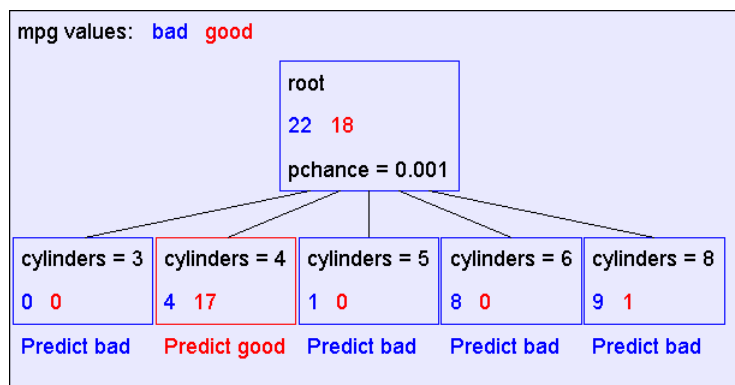
mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europa
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europa
bad	5	medium	medium	medium	medium	75to78	europa

## Find the First Split

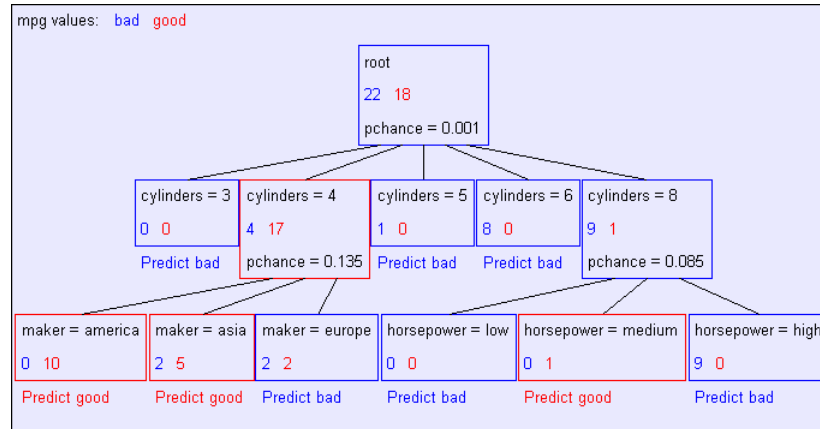
- Look at information gain for each attribute
- Note that each attribute is correlated with the target!
- What do we split on?



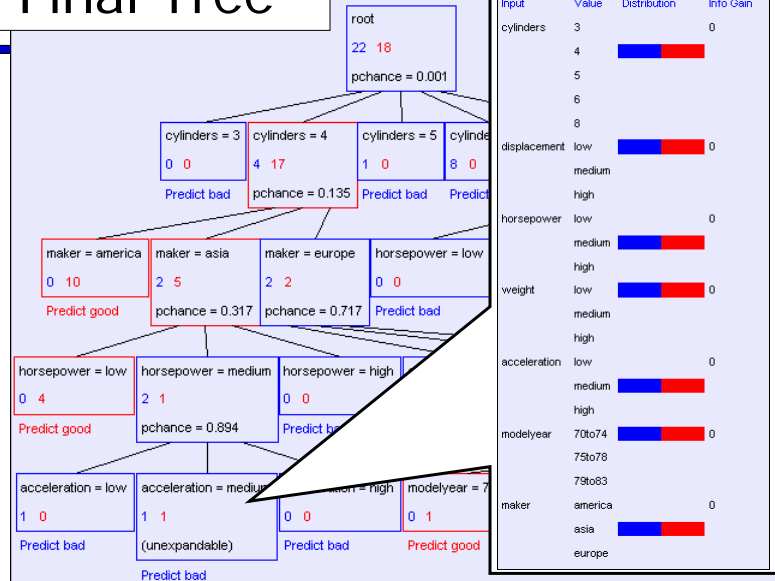
## Result: Decision Stump



## Second Level



## Final Tree



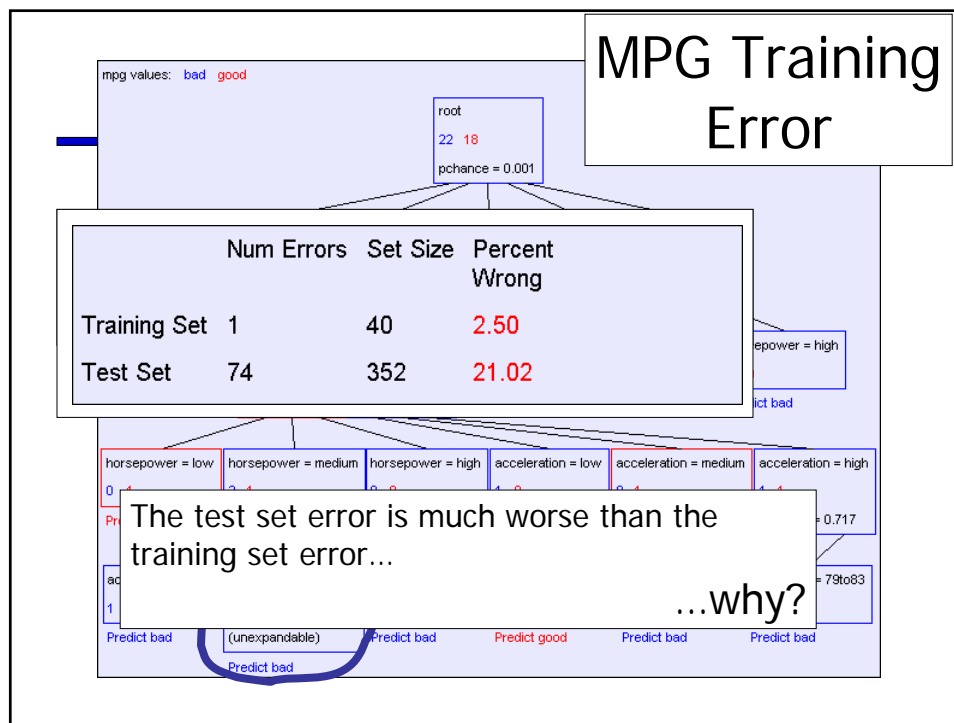
## Reminder: Overfitting

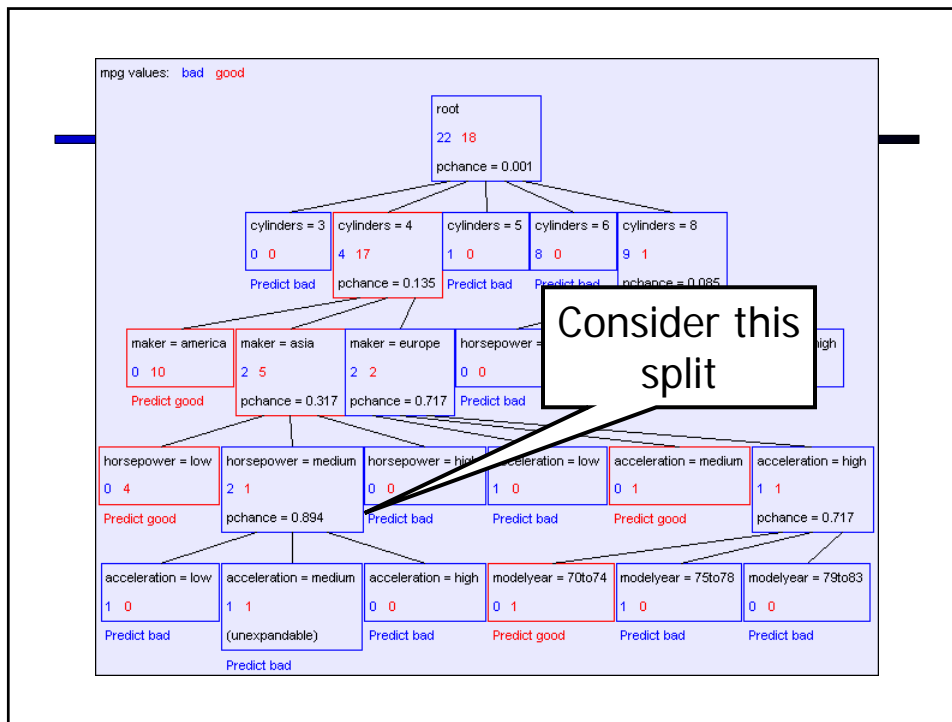
- Overfitting:

- When you stop modeling the patterns in the training data (which generalize)
- And start modeling the noise (which doesn't)

- We had this before:

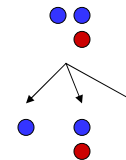
- Naïve Bayes: needed to smooth
- Perceptron: didn't really say what to do about it (stay tuned!)





## Significance of a Split

- Starting with:
  - Three cars with 4 cylinders, from Asia, with medium HP
  - 2 bad MPG
  - 1 good MPG
- What do we expect from a three-way split?
  - Maybe each example in its own subset?
  - Maybe just what we saw in the last slide?
- Probably shouldn't split if the counts are so small they could be due to chance
- A chi-squared test can tell us how likely it is that deviations from a perfect split are due to chance (details in the book)
- Each split will have a **significance value**,  $p_{\text{CHANCE}}$



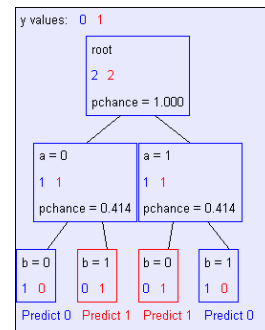
# Keeping it General

## Pruning:

- Build the full decision tree
- Begin at the bottom of the tree
- Delete splits in which  $P_{\text{CHANCE}} > \text{Max}P_{\text{CHANCE}}$
- Continue working upward until there are no more prunable nodes
- Note: some chance nodes may not get pruned because they were “redeemed” later

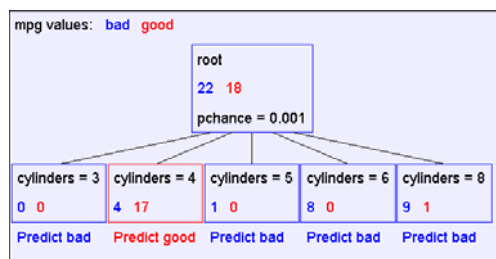
$$y = a \text{ XOR } b$$

a	b	y
0	0	0
0	1	1
1	0	1
1	1	0



# Pruning example

- With  $\text{Max}P_{\text{CHANCE}} = 0.1$ :

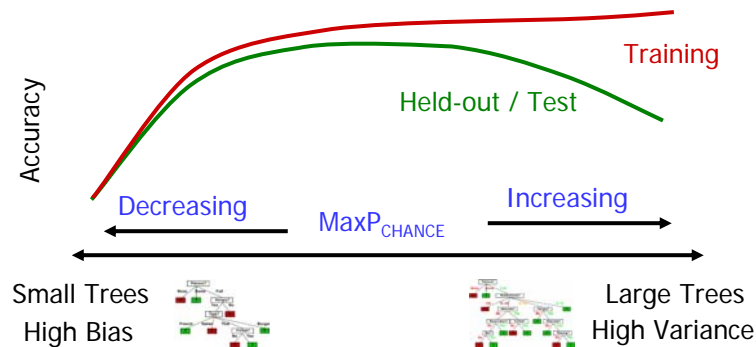


Note the improved test set accuracy compared with the unpruned tree

	Num Errors	Set Size	Percent Wrong
Training Set	5	40	12.50
Test Set	56	352	15.91

# Regularization

- $\text{MaxP}_{\text{CHANCE}}$  is a regularization parameter
- Generally, set it using held-out data (as usual)



## Two Ways of Controlling Overfitting

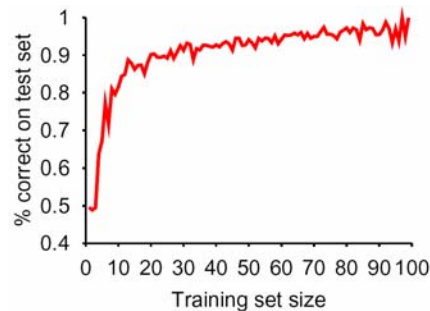
- Limit the hypothesis space
  - E.g. limit the max depth of trees
  - Easier to analyze (coming up)
- Regularize the hypothesis selection
  - E.g. chance cutoff
  - Disprefer most of the hypotheses unless data is clear
  - Usually done in practice

# Learning Curves

---

- Another important trend:

- More data is better!
- The same learner will generally do better with more data
- (Except for cases where the target is absurdly simple)



# Summary

---

- Formalization of learning

- Target function
- Hypothesis space
- Generalization

- Decision Trees

- Can encode any function
- Top-down learning (not perfect!)
- Information gain
- Bottom-up pruning to prevent overfitting