CSE 446: Decision Trees
Winter 2012

Slides adapted from Carlos Guestrin and Andrew Moore by Luke Zettlemoyer; tweaked by Dan Weld

Logistics

- PS 1 – due in two weeks
  - Did we already suggest that you … start early
- See website for reading

Why is Learning Possible?

Experience alone never justifies any conclusion about any unseen instance.

Learning occurs when
PREJUDICE meets DATA!

Some Typical Biases

- Occam's razor
  - ***"It is needless to do more when less will suffice"***
  - ***– William of Occam,***
    died 1349 of the Black plague
- MDL – Minimum description length
- Concepts can be approximated by
  - … **conjunctions** of predicates
    … by **linear** functions
    … by **short** decision trees

Overview of Learning

What is Being Learned?

Type of Supervision
(eg, Experience, Feedback)

	Labeled Examples	Reward	Nothing
Discrete Function	Classification		Clustering
Continuous Function	Regression		
Policy	Apprenticeship Learning	Reinforcement Learning	

A learning problem: predict fuel efficiency

- 40 Records
- Discrete data (for now)
- Predict MPG

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	europe
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	europe
bad	5	medium	medium	medium	medium	75to78	europe

From the UCI repository (thanks to Ross Quinlan)

Past Insight

Any ML problem may be

cast as the problem of

FUNCTION APPROXIMATION

A learning problem: predict fuel efficiency

- 40 Records
- Discrete data (for now)
- Predict MPG

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	europe
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	europe
bad	5	medium	medium	medium	medium	75to78	europe

Y X

Need to find "Hypothesis": $f : X \rightarrow Y$

How Represent Function?

f (| cylinders | displacement | horsepower | weight | acceleration | modelyear | maker |) → | mpg |

4	low	low	low	high	75to78	asia		good

Conjunctions in Propositional Logic?

maker=asia ∧ weight=low

Need to find "Hypothesis": $f : X \rightarrow Y$

Restricted Hypothesis Space

- Many possible representations
- Natural choice: ***conjunction*** of attribute constraints
- For each attribute:
  - Constrain to a specific value: eg **maker=asia**
  - Don't care: **?**
- For example

maker cyl displace weight accel ....

asia ? ? low ?

Represents **maker=asia ∧ weight=low**

Consistency

- Say an "example is consistent with a hypothesis" when the example *logically satisfies* the hypothesis
- Hypothesis: **maker=asia ∧ weight=low**

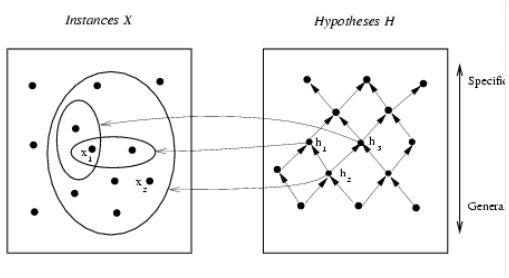
maker cyl displace weight accel ....

asia ? ? low ?

- Examples:

asia	5	low	low	low	...
usa	4	low	low	low	...

Ordering on Hypothesis Space



x_1	asia	5	low	low	low
x_2	usa	4	med	med	med

h1: maker=asia ∧ accel=low

h2: maker=asia

h3: maker=asia ∧ weight=low

Version Space Algorithm

Ok, so how does it perform?

13

How Represent Function?

f (	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
4	low	low	low	high	75to78	asia

) → | mpg |
|---|
| good |

General Propositional Logic?

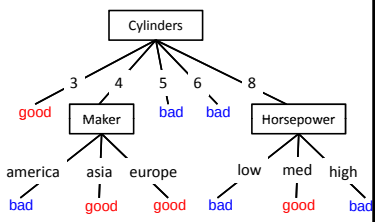
maker=asia ∨ weight=low

Need to find "Hypothesis": $f : X \rightarrow Y$

Hypotheses: decision trees $f : X \rightarrow Y$

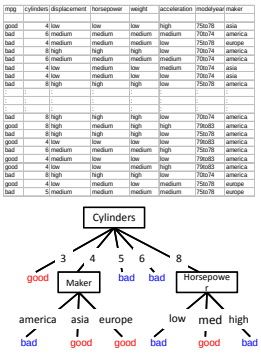
- Each internal node tests an attribute x_i
- Each branch assigns an attribute value $x_i=v$
- Each leaf assigns a class y
- To classify input x?

  traverse the tree from root to leaf, output the labeled y



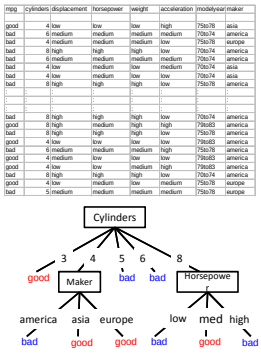
Hypothesis space

- How many possible hypotheses?

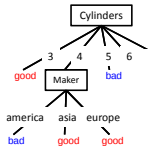


Hypothesis space

- How many possible hypotheses?
- What functions can be represented?



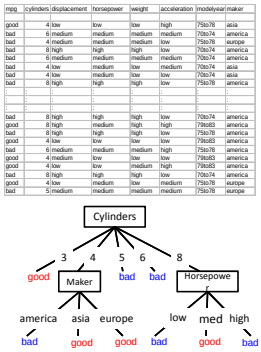
What functions can be represented?



cyl=3 ∨ (cyl=4 ∧ (maker=asia ∨ maker=europe)) ∨ …

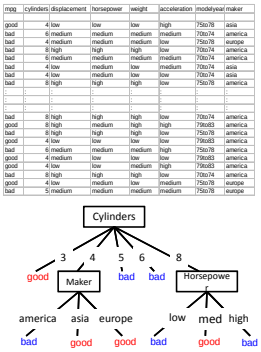
Hypothesis space

- How many possible hypotheses?
- What functions can be represented?
- How many will be consistent with a given dataset?



Hypothesis space

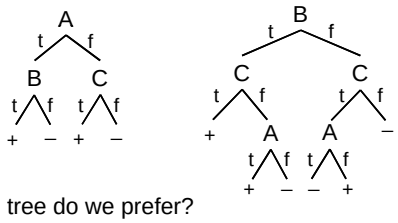
- How many possible hypotheses?
- What functions can be represented?
- How many will be consistent with a given dataset?
- How will we choose the best one?



Are all decision trees equal?

- Many trees can represent the same concept
- But, not all trees will have the same size!

  e.g., $\phi = (A \wedge B) \vee (\neg A \wedge C)$



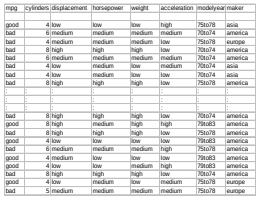
- Which tree do we prefer?

Learning decision trees is hard!!!

- Finding the simplest (smallest) decision tree is an NP-complete problem [Hyafil & Rivest '76]
- Resort to a ***greedy*** heuristic:
  - Start from empty decision tree
  - Split on **next best attribute (feature)**
  - Recurse

What is the Simplest Tree?

predict mpg=bad

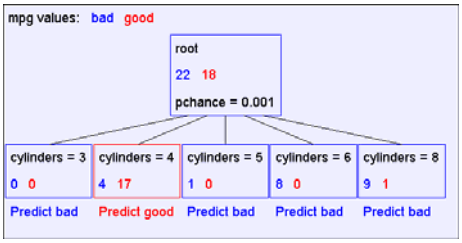


Is this a good tree?

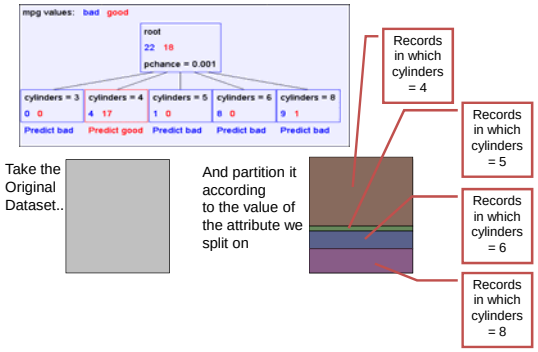
[22+, 18-] ← Means:
correct on 22 examples
incorrect on 18 examples

Improving Our Tree

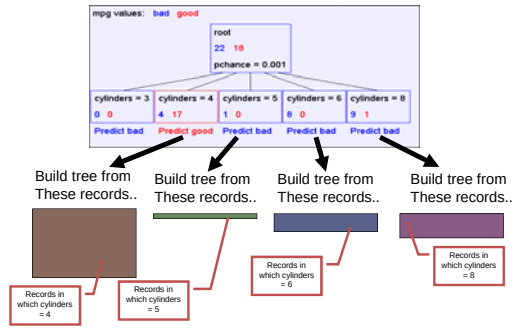
predict mpg=bad



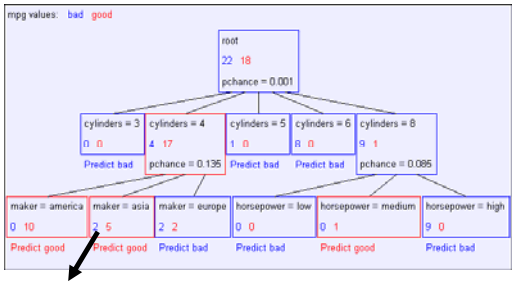
Recursive Step



Recursive Step



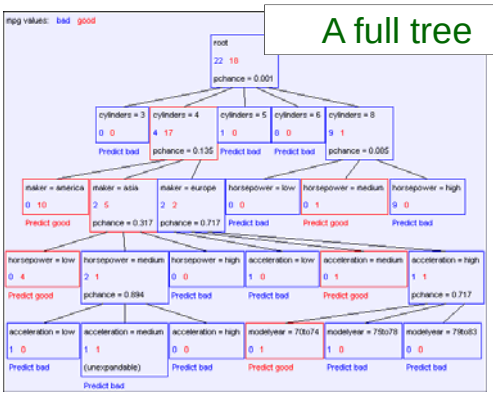
Second level of tree



Recursively build a tree from the seven records in which there are four cylinders and the maker was based in Asia

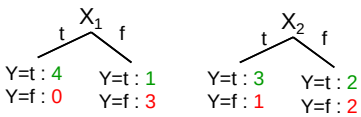
(Similar recursion in the other cases)

A full tree



Splitting: choosing a good attribute

Would we prefer to split on X_1 or X_2?



X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

Idea: use counts at leaves to define probability distributions so we can measure uncertainty!

Measuring uncertainty

- Good split if we are more certain about classification after split
  - Deterministic good (all true or all false)
  - Uniform distribution? Bad
  - What about distributions in between?

P(Y=A) = 1/2	P(Y=B) = 1/4	P(Y=C) = 1/8	P(Y=D) = 1/8

P(Y=A) = 1/3	P(Y=B) = 1/4	P(Y=C) = 1/4	P(Y=D) = 1/6

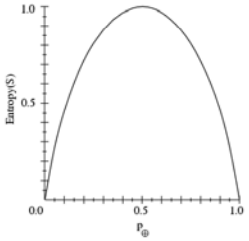
Entropy

Entropy $H(Y)$ of a random variable Y

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$

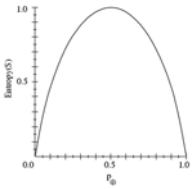
More uncertainty, more entropy!

Information Theory interpretation:
$H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)



Entropy Example

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$



P(Y=t) = 5/6
P(Y=f) = 1/6

H(Y) = - 5/6 $\log_2$ 5/6 - 1/6 $\log_2$ 1/6
= 0.65

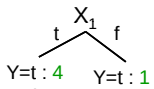
X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F

Conditional Entropy

Conditional Entropy $H(Y|X)$ of a random variable Y conditioned on a random variable X

$$H(Y \mid X) = -\sum_{j=1}^{v} P(X = x_j) \sum_{i=1}^{k} P(Y = y_i \mid X = x_j) \log_2 P(Y = y_i \mid X = x_j)$$

Example:



Y=t : 4 Y=t : 1
Y=f : 0 Y=f : 1

P(X_1=t) = 4/6
P(X_1=f) = 2/6

H(Y|X_1) = - 4/6 (1 $\log_2$ 1 + 0 $\log_2$ 0)
- 2/6 (1/2 $\log_2$ 1/2 + 1/2 $\log_2$ 1/2)
= 2/6
= 0.33

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F

Information Gain

Advantage of attribute – decrease in entropy (uncertainty) after splitting

$$IG(X) = H(Y) - H(Y \mid X)$$

In our running example:

IG(X_1) = H(Y) – H(Y|X_1)
= 0.65 – 0.33

IG(X_1) > 0 → we prefer the split!

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F

Learning Decision Trees

- Start from empty decision tree
- Split on **next best attribute (feature)**
  - Use, for example, information gain to select attribute:

$$\arg\max_i IG(X_i) = \arg\max_i H(Y) - H(Y \mid X_i)$$

- Recurse

Acknowledgements

- Some of the material in the decision trees presentation is courtesy of Andrew Moore, from his excellent collection of ML tutorials:
  - http://www.cs.cmu.edu/~awm/tutorials