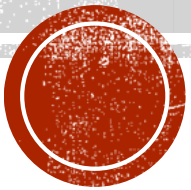
DECISION TREES



CS 412 Introduction to Machine Learning

Prof. Zheleva

January 18, 2018

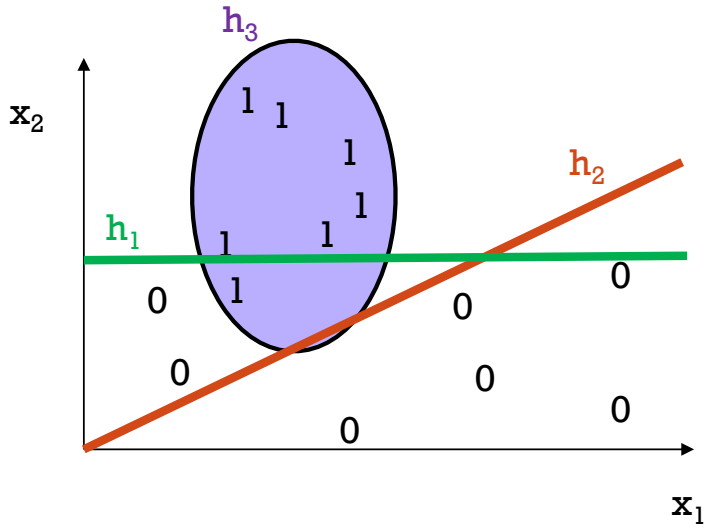
Reading Assignment: CIML: 1

LAST LECTURE: MACHINE LEARNING INTRO

- What does learning by example mean?
  - Classification task
  - Learning requires examples
    - Generalization vs. memorization
  - Formalizing the learning problem
    - Function approximation
    - Learning as minimizing the expected loss

SUPERVISED LEARNING SETTING

- Problem setting
  - **X** – set of possible instances
  - Unknown target function f: **X** $\rightarrow$ Y
    - Classification: Y is discrete valued
    - Regression: Y is real-valued
    - Density estimation: probability of each $y \in Y$
  - Set of function hypotheses H={h | h: **X** $\rightarrow$ Y}



- Input
  - Training examples $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots (\boldsymbol{x}^{(N)}, y^{(N)})\}$ of unknown distribution

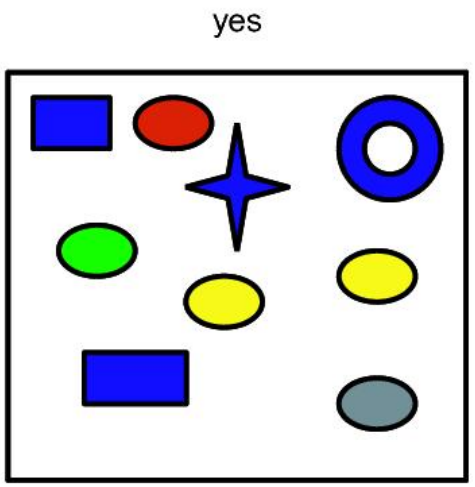
- Output
  - Hypothesis $h \in$ H that best approximates target function f

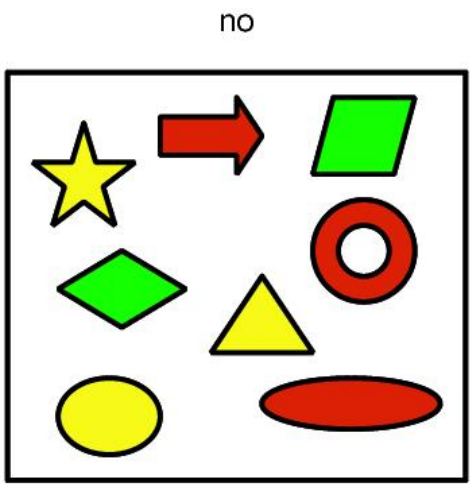
TODAY: DECISION TREES

- **What is a decision tree?**
- How to learn a decision tree from data?
- What is the inductive bias?
- Generalization

Toy dataset and classification problem

Are these objects from outer space?





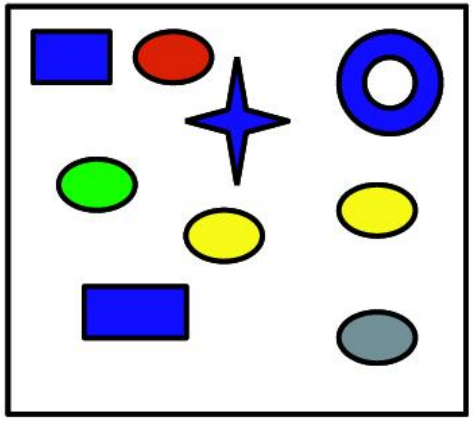
D features (attributes)

N cases

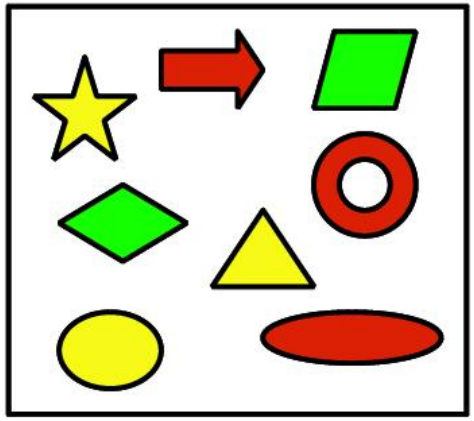
Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Label
1
1
0

yes



no



20 Questions game

Red	Ellipse	20.7	0

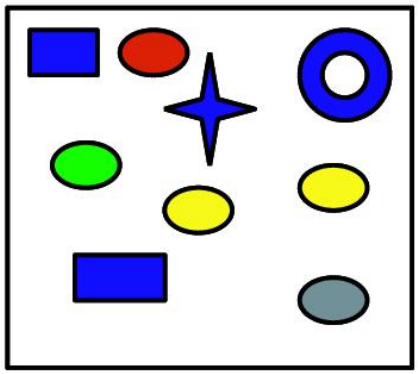
DECISION TREES

- "20 questions game for each possible outcome"

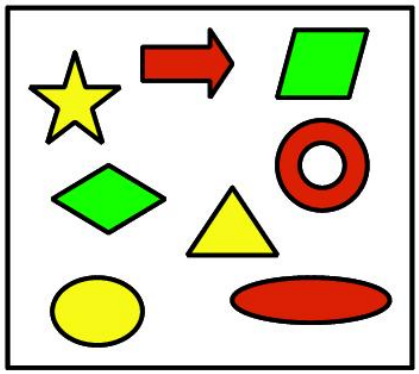
- Representation
  - **Nodes:** test the value of feature x_j
  - **Branches**: correspond to values
  - **Leafs**: provide the class (prediction)
    - Or a probability distribution over classifications

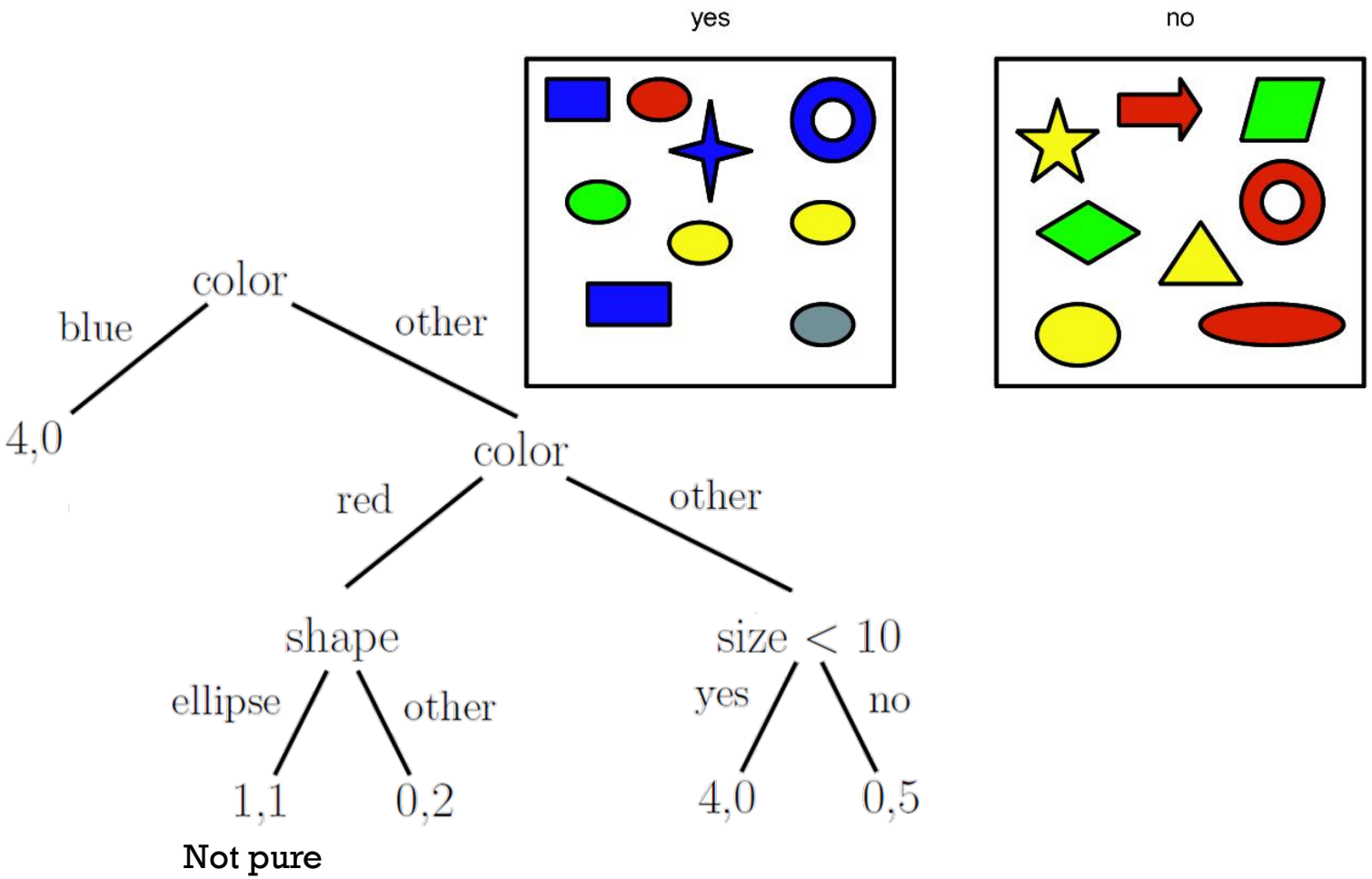
- Decision trees represent functions that map examples in X to classes in Y
  - f: <Color,Shape,Size> -> Yes/No?

yes

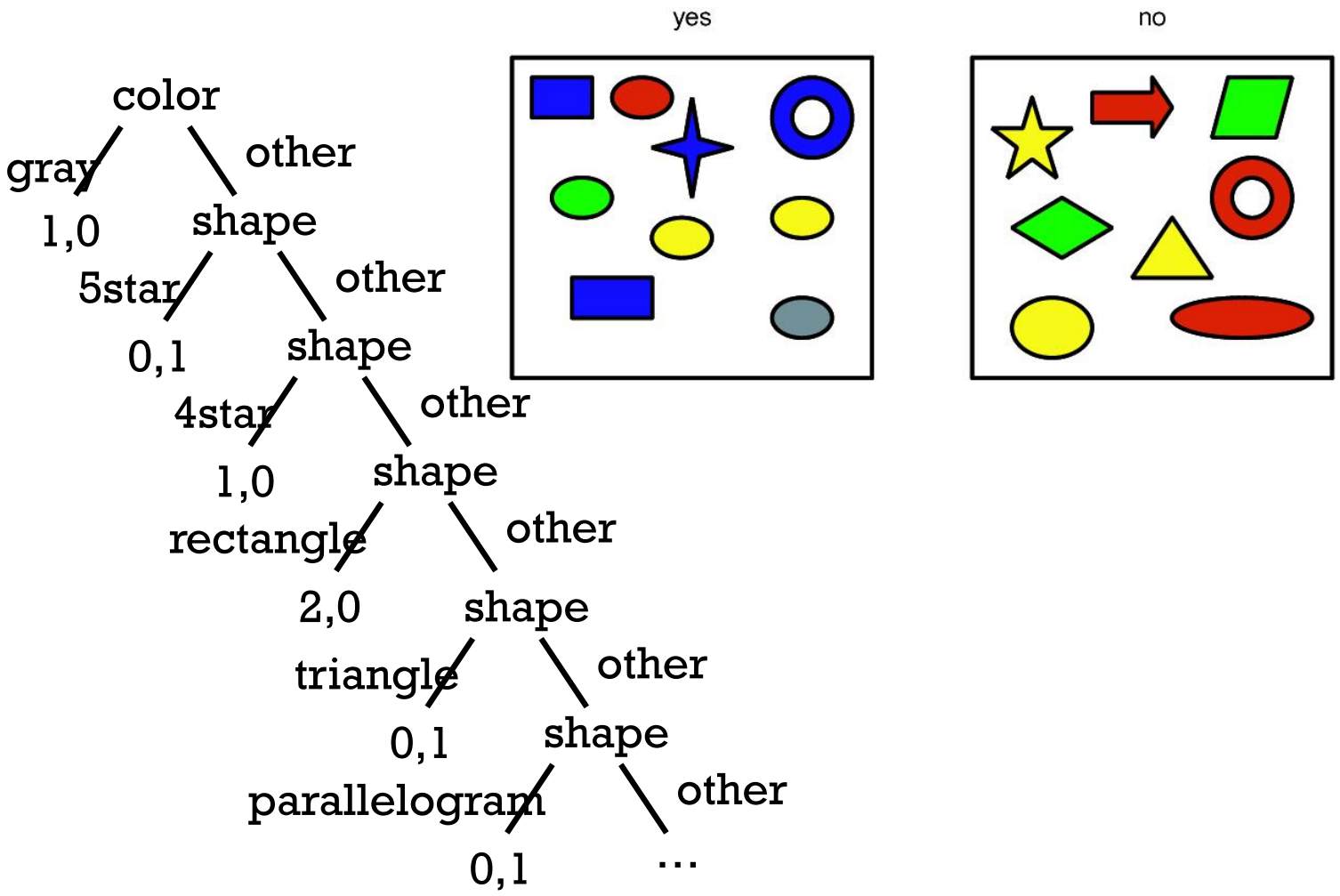


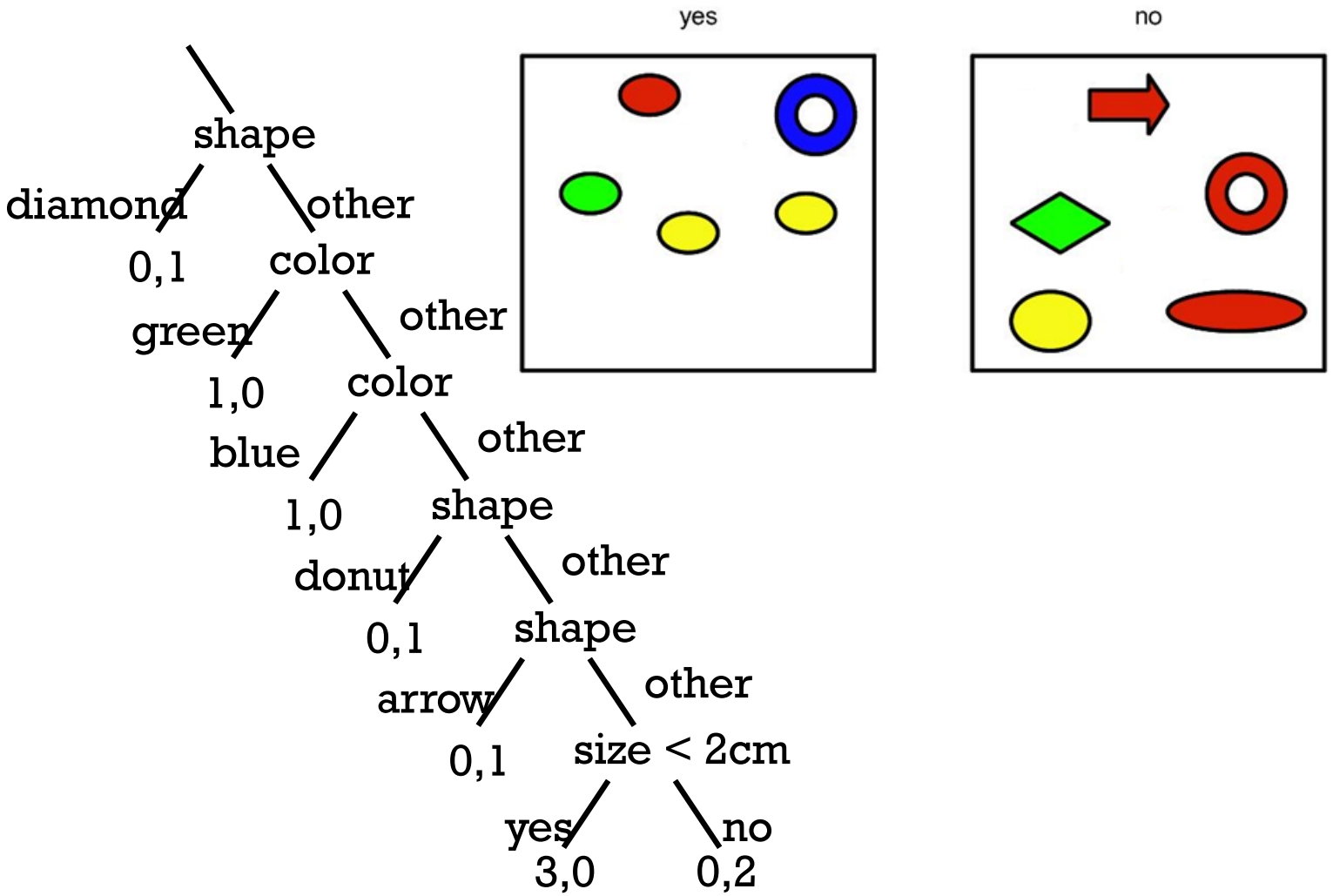
no




yes
no
color
blue
other
4,0
color
red
other
shape
size < 10
ellipse
other
yes
no
1,1
0,2
4,0
0,5

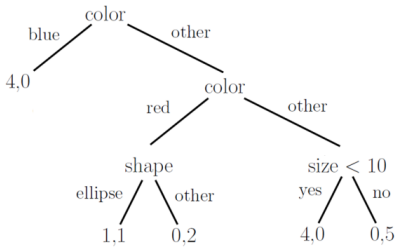
Not pure


yes
no
color
gray
other
1,0
shape
5star
other
0,1
shape
4star
other
1,0
shape
rectangle
other
2,0
shape
triangle
other
0,1
shape
parallelogram
other
0,1
…


yes
no
shape
diamond
other
0,1
color
green
other
1,0
color
blue
other
1,0
shape
donut
other
0,1
shape
arrow
other
0,1
size < 2cm
yes
no
3,0
0,2

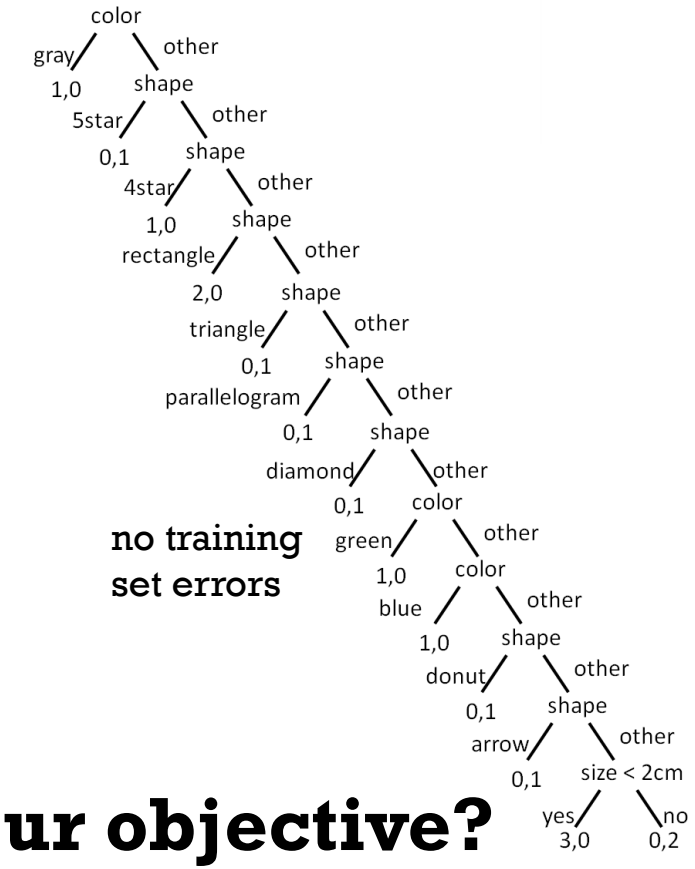
EXERCISE

- How would you represent the following Boolean functions with decision trees?
  - AND
  - OR
  - XOR
  - (A OR B) AND (C OR not(D))



1 training set error

or



no training set errors

What is our objective?

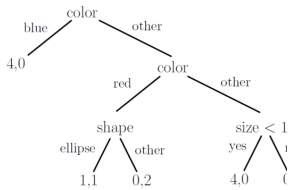
TODAY: DECISION TREES

- What is a decision tree?
- **How to learn a decision tree from data?**
- What is the inductive bias?
- Generalization

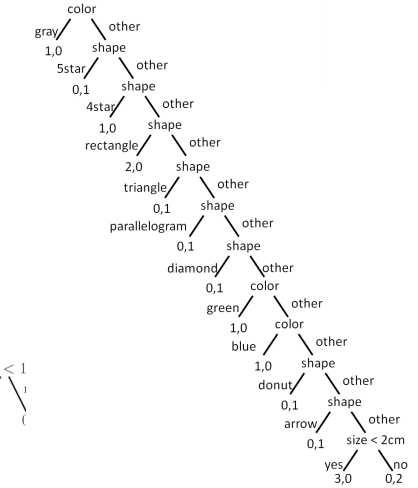
SUPERVISED LEARNING SETTING

- Problem setting
  - **X** – set of possible instances
    - Each instance $\boldsymbol{x} \in X$ is a feature vector $\boldsymbol{x} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_D}]$
  - Unknown target function f: **X** $\rightarrow$ Y
    - Classification: Y is discrete valued
  - Set of function hypotheses H={h | h: **X** $\rightarrow$ Y}
    - Each hypothesis h is a decision tree

$\boldsymbol{h_1}$:



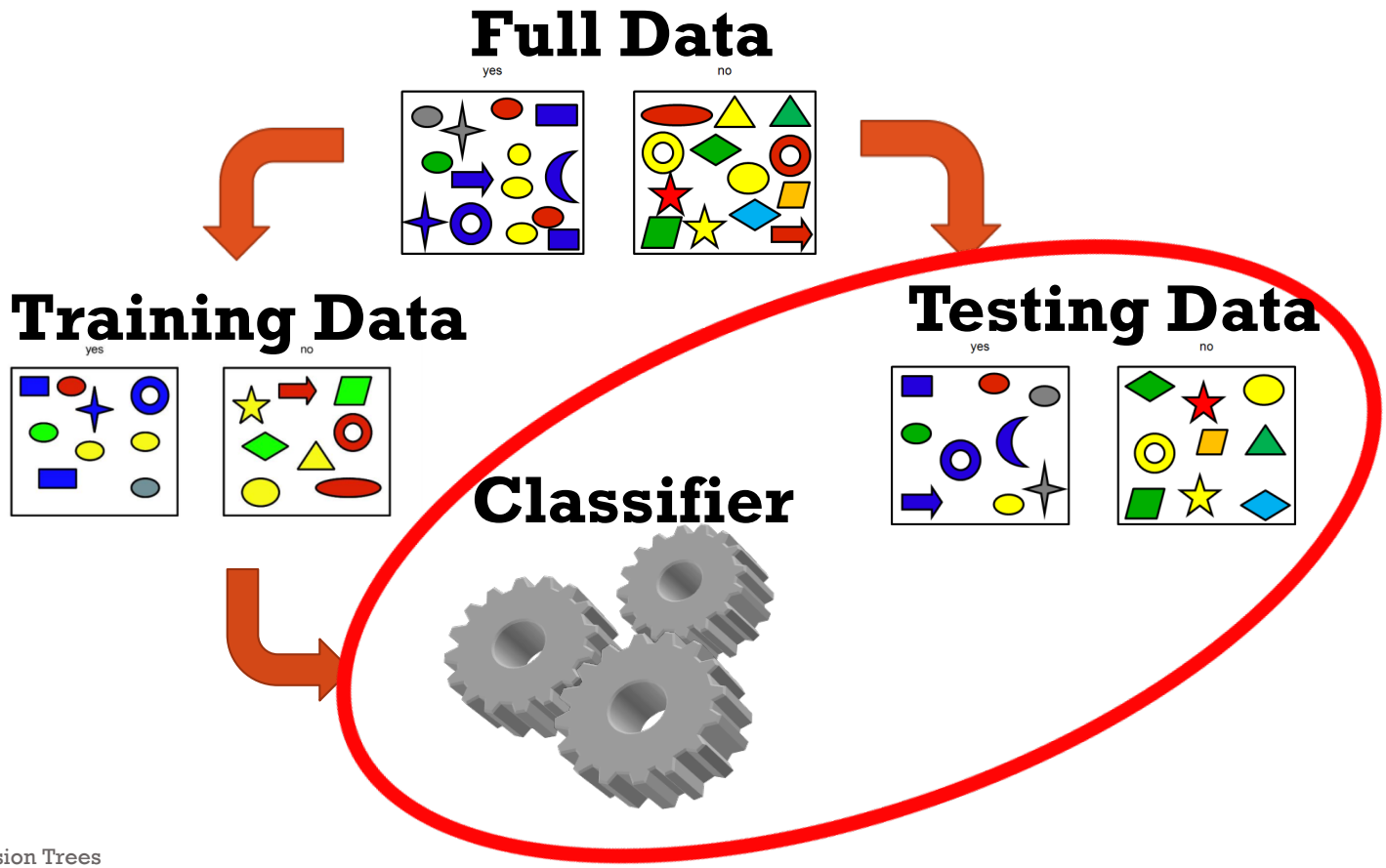
$\boldsymbol{h_2}$:



- Input
  - Training examples $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots (\boldsymbol{x}^{(N)}, y^{(N)})\}$ of unknown distribution

- Output
  - Hypothesis $h \in H$ that best approximates target function f

SUPERVISED LEARNING



SUPERVISED LEARNING

- Main assumption:
  - **Training Data** and **Testing Data** are drawn from the same distribution.
  - Also known as: **identically and independently distributed** (**iid** or **IID**)

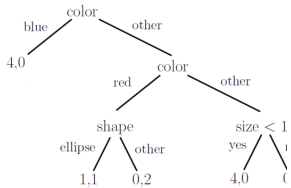
DECISION TREE LEARNING

- <u>Goal:</u> find the tree $h \in H$
  - That minimizes training error
  - Or maximizes training accuracy

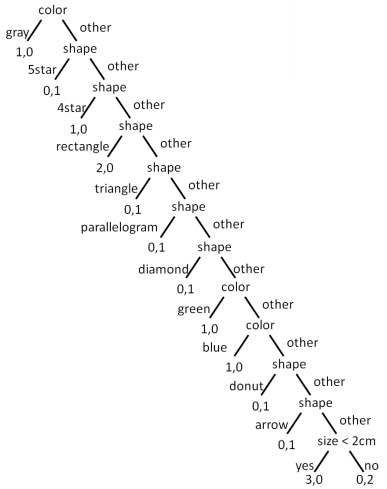
- H is too large for exhaustive search
  - Finding the best tree is NP-Hard!

- Instead, use a heuristic search algorithm which
  - Picks questions to ask, in order
  - Such that classification accuracy is maximized

h_1:



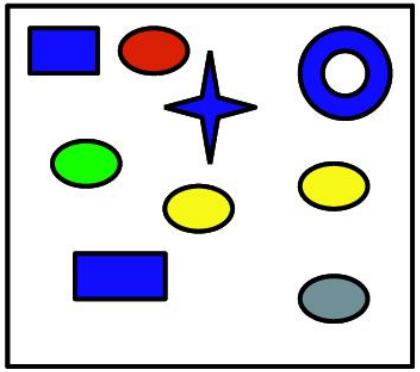
h_2:



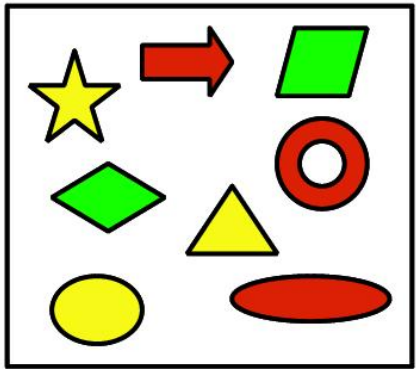
HOW TO SELECT THE BEST FEATURE

- A good feature is a feature that lets us make correct classification decisions
- We will use Score(D_1, D_2) measuring "goodness" of splits of the data D into D_1 and D_2 datasets
  - For example, select features based on error rate
- Let's try it on our toy dataset

yes

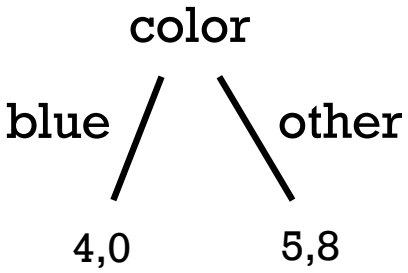


no

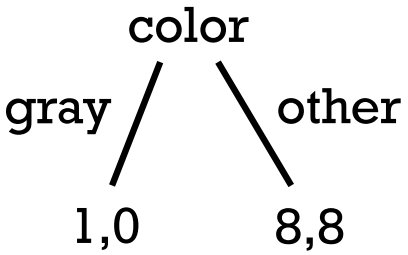


f: <Color,Shape,Size> -> Yes/No?

Consider these two possible tree splits:



Error rate: 5/17



Error rate: 8/17

To classify: use majority label at the leaf

TOP-DOWN INDUCTION OF DECISION TREES

CurrentNode = Root

DTtrain(examples for CurrentNode,features at CurrentNode):

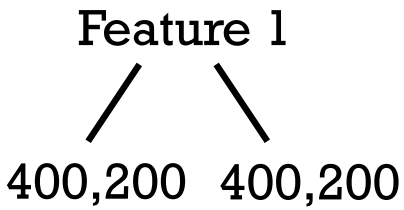
1. Find F, the decision feature for next node with best Score(D_1, D_2)
2. For each value of F, create new descendant of node
3. Sort training examples to leaf nodes
4. If training examples perfectly classified

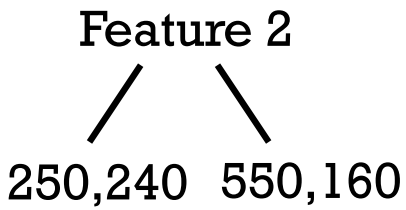
   Stop

   Else

   Recursively apply DTtrain over new leaf nodes

ACCURACY SCORE PITFALL





Error rate: 400/1200

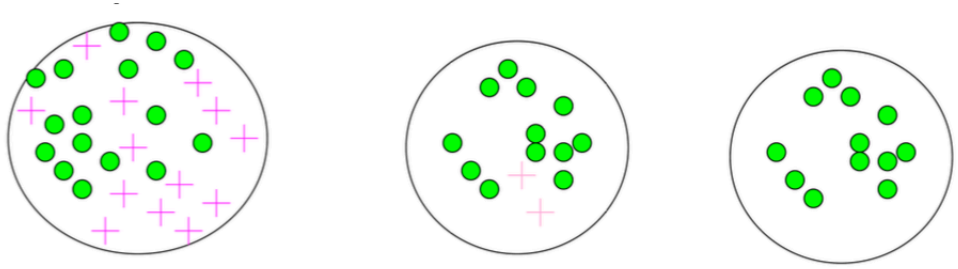
Error rate: 400/1200

Both have the same error rate!!!

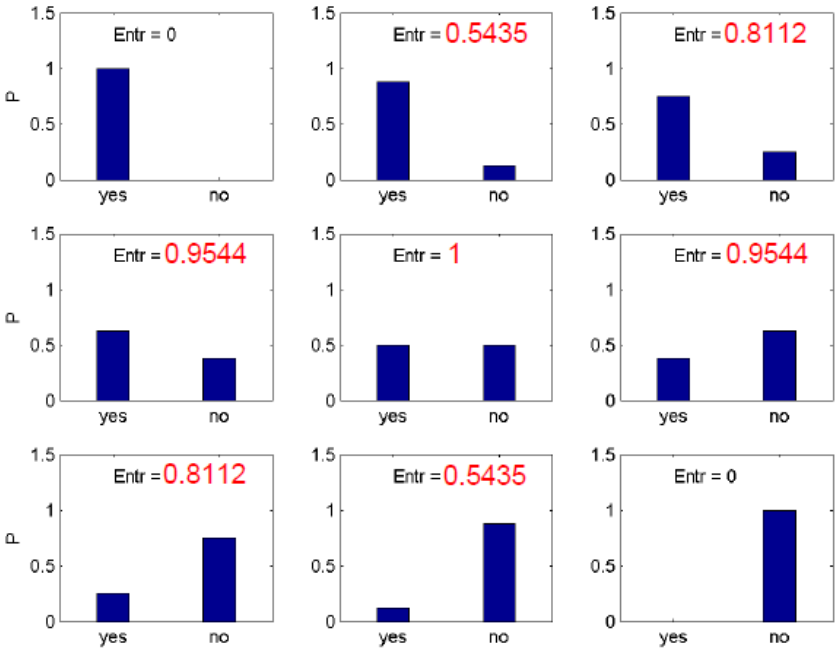
Which is "progressing more" towards a lower error?

ANOTHER SCORING FUNCTION: ENTROPY

- Entropy measures sample impurity or (degree of) uncertainty



- Used in the ID3 algorithm [Quinlan, 1986]
  - Pick feature with the smallest entropy to split the examples



The entropy is maximal when all possibilities are equally likely.

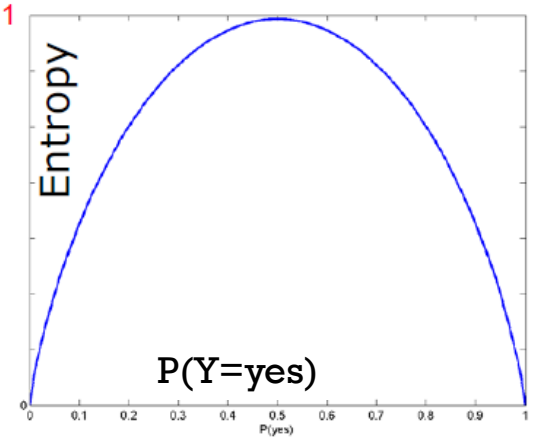
The goal of the decision tree is to decrease the entropy in each node.

Entropy is zero in a pure "yes" node (or pure "no" node).

Entropy is a measure of "uncertainty" of a random variable.

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

n – number of possible values for the random variable



ENTROPY

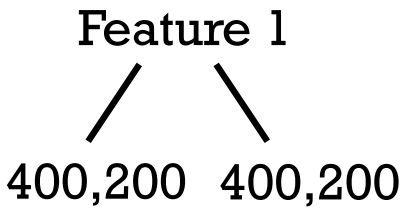
- Where does the formula come from?

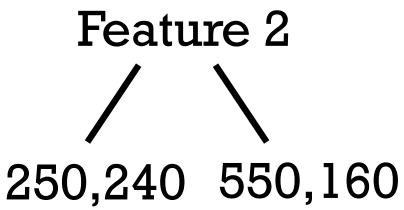
$$\mathrm{H(D)} = \sum_{i=1}^{n} -P(Y = i)\log_2 P(Y = i)$$

- Information theory
  - Entropy is the expected number of bits needed to encode a randomly drawn value of Y
  - Most efficient possible code assigns $-\log_2 P(Y = i)$ bits to encode the message $\mathrm{Y} = i$
  - Lower probability events carry higher value

- Roll of a 4 sided fair die. What is its entropy?

REVISIT: ACCURACY SCORE PITFALL





Error rate: 400/1200
Entropy: 0.92

Error rate: 400/1200
Entropy: 0.86

ALTERNATIVE SCORE: INFORMATION GAIN*

- Information gain is the measure of the difference in entropy before and after a split on a given attribute A

$$\mathrm{IG}(A, D) = H(D) - \sum_{D_i \in \boldsymbol{D}} p(D_i) H(D_i)$$

  - H(D) – entropy of dataset D
  - $\boldsymbol{D}$ - subsets created from splitting D by attribute A
  - p($\mathrm{D_i}$) – $|\mathrm{D_i}|/|D|$

- Used in C4.5 algorithm by Quinlan (1993)

ALTERNATIVE SCORE: GINI IMPURITY

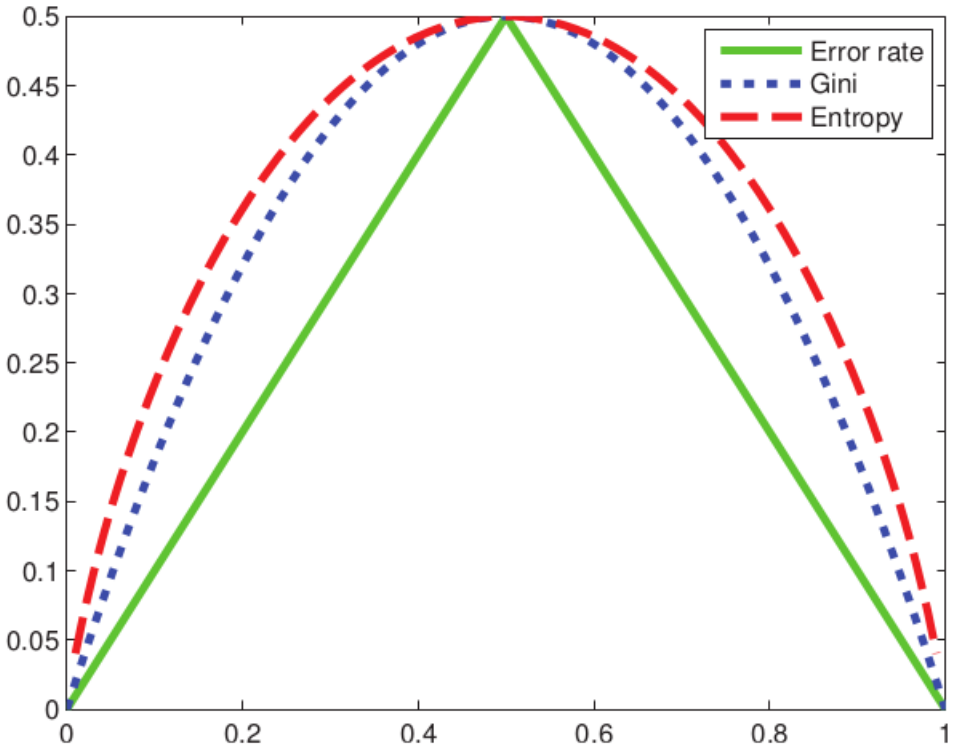
- Measure of how often a randomly chosen element from the set would be incorrectly labeled if randomly labeled according to the distribution of labels in the set

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

  - J is the set of classes
  - p_i is the probability of an item with label i being chosen
  - $\sum_{k \neq i} \; p_k$ is the probability of a mistake in classifying item i

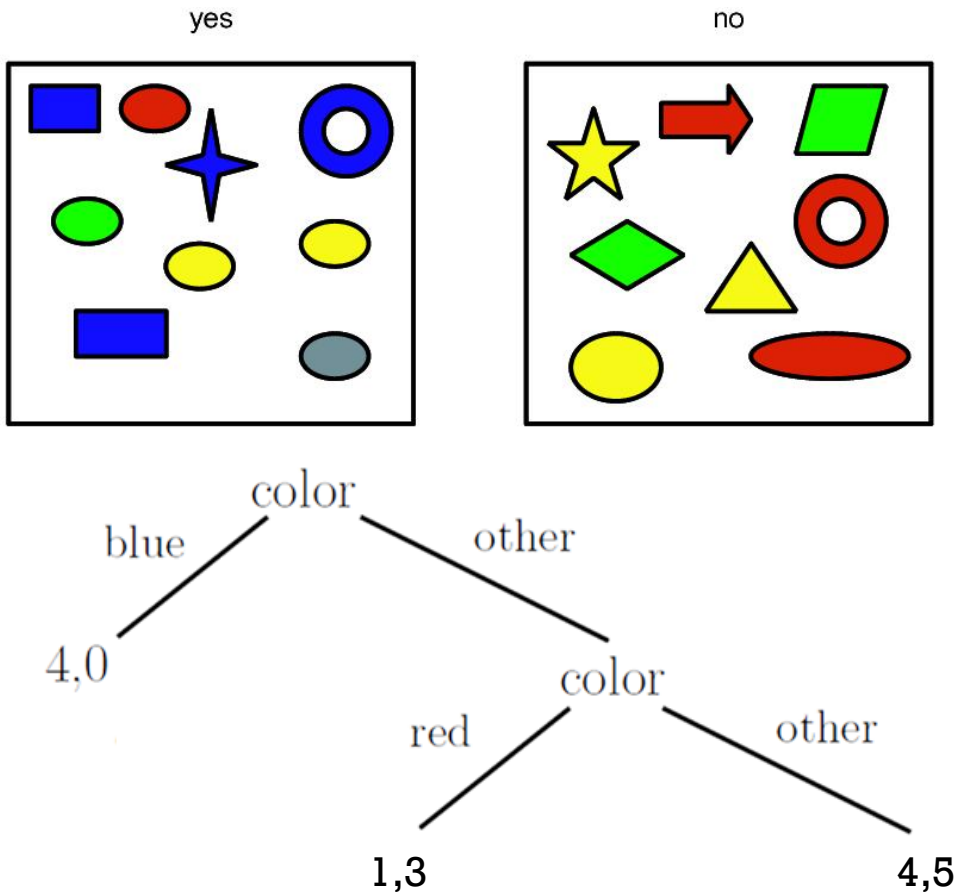
- Used by the CART algorithm of Breiman et al. (1984)

IMPURITY MEASURES



P (X=1) binary

Entropy rescaled


yes
no
color
blue
other
4,0
color
red
other
1,3
4,5

DECISION TREE VISUALIZATION

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

TODAY: DECISION TREES

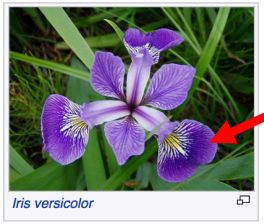
- What is a decision tree?
- How to learn a decision tree from data?
- **What is the inductive bias?**
- Generalization

OVERFITTING

High accuracy on training data, low accuracy on testing data

Let's look at another example

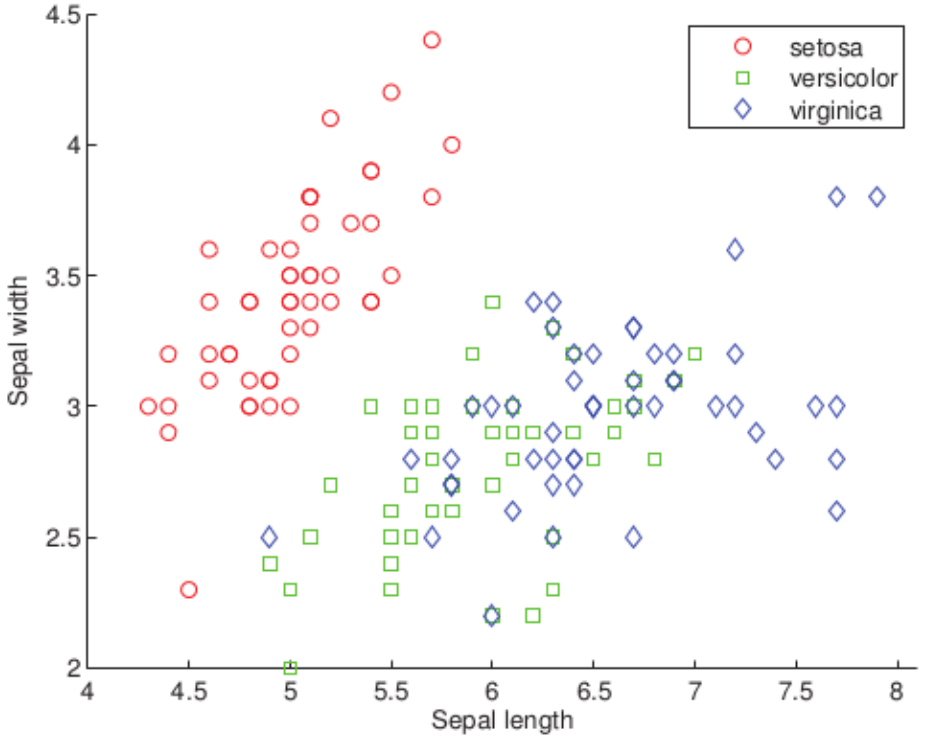
Iris setosa

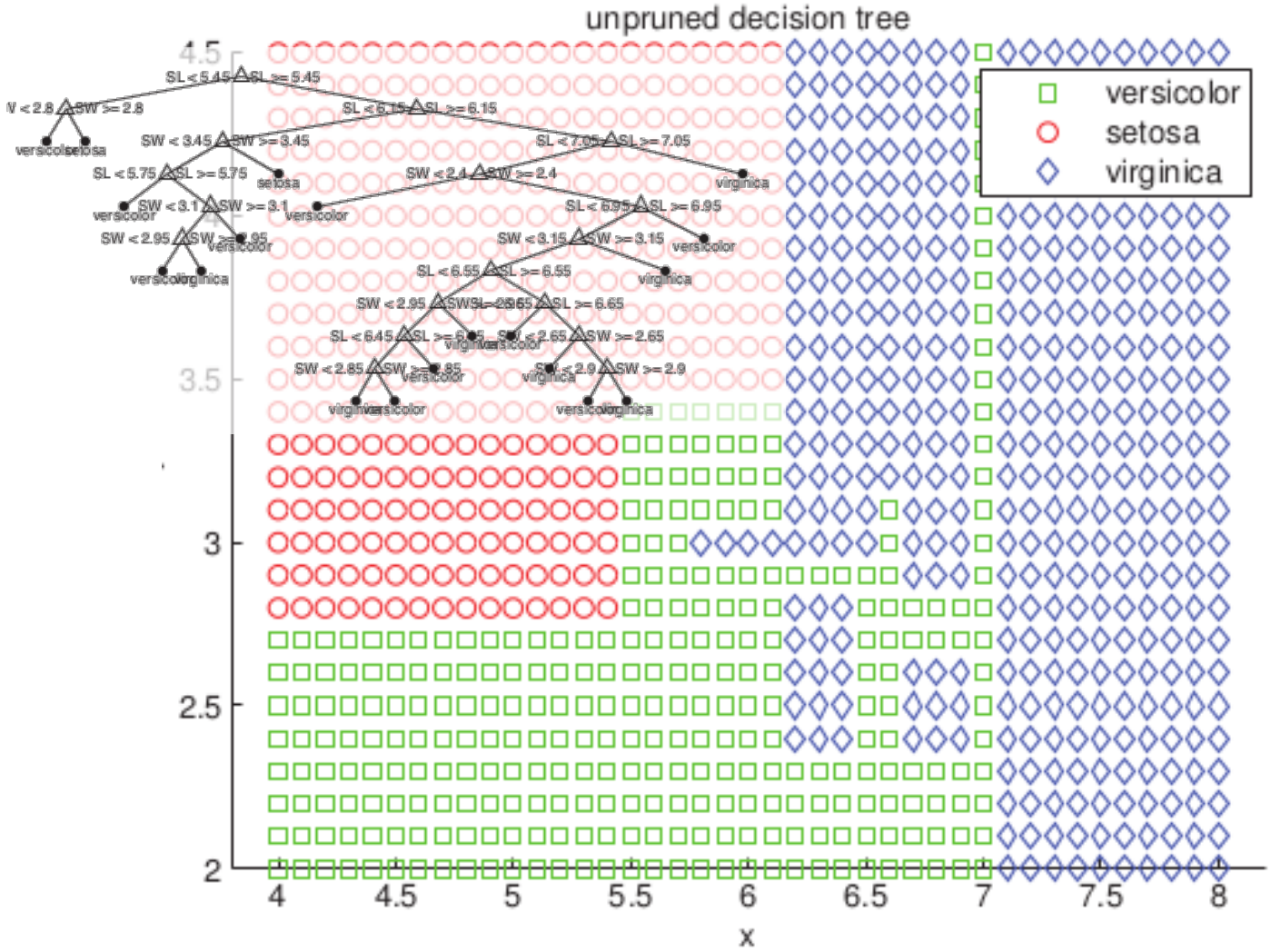


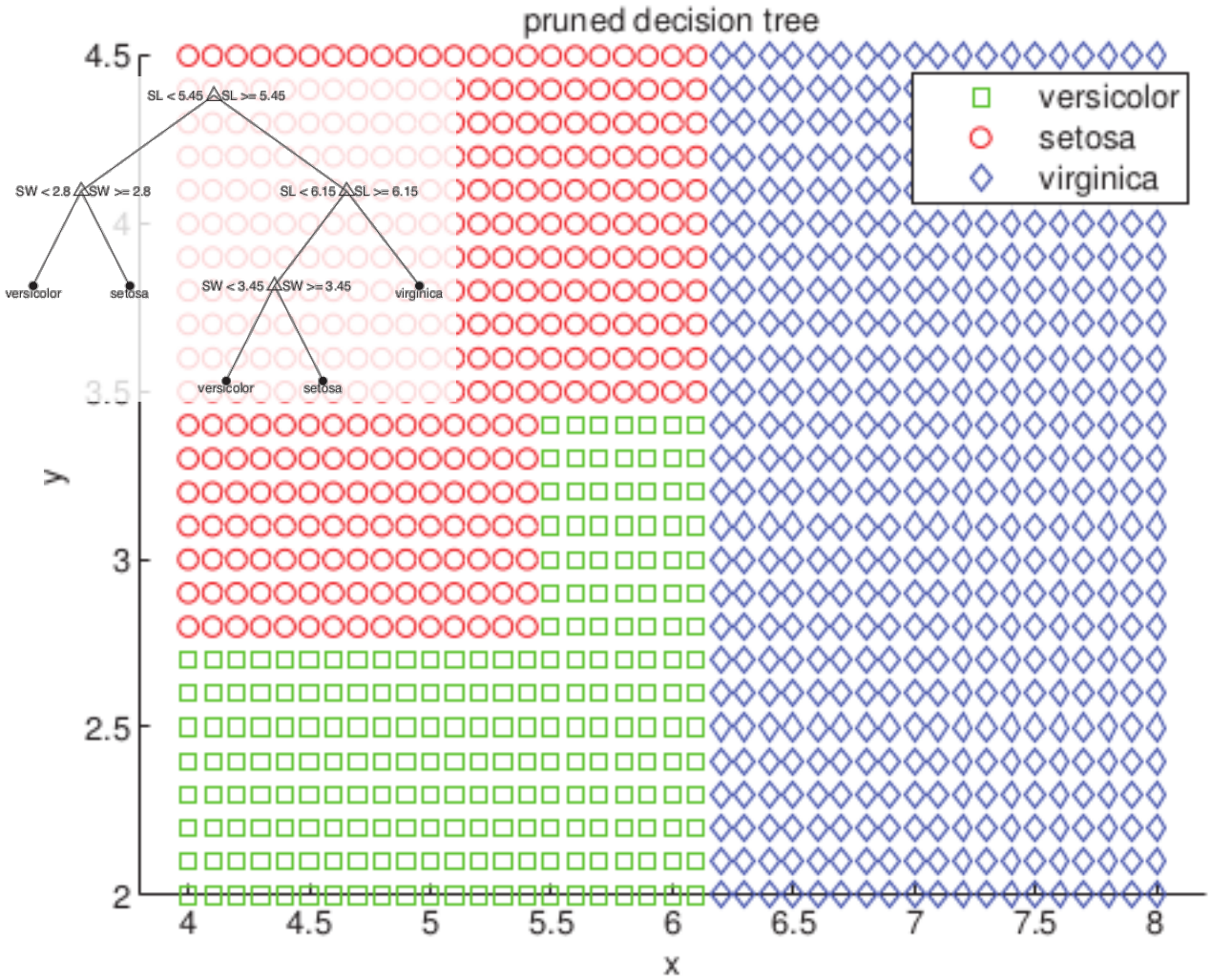
Iris versicolor

Sepal

Iris virginica




unpruned decision tree
versicolor
setosa
virginica
SL < 5.45
SL >= 5.45
SW < 2.8
SW >= 2.8
versicolor
setosa
SL < 6.15
SL >= 6.15
SW < 3.45
SW >= 3.45
SL < 7.05
SL >= 7.05
SL < 5.75
SL >= 5.75
setosa
SW < 2.4
SW >= 2.4
virginica
SW < 3.1
SW >= 3.1
versicolor
versicolor
SL < 6.95
SL >= 6.95
SW < 2.95
SW >= 2.95
versicolor
SW < 3.15
SW >= 3.15
versicolor
versicolor
virginica
SL < 6.55
SL >= 6.55
virginica
SW < 2.95
SW >= 2.95
SL < 6.65
SL >= 6.65
SL < 6.45
SL >= 6.45
versicolor
SL < 2.65
SW >= 2.65
SW < 2.85
SW >= 2.85
versicolor
SL < 2.9
SW >= 2.9
virginica
virginica
versicolor
versicolor
virginica
4.5
3.5
3
2.5
2
4
4.5
5
5.5
6
6.5
7
7.5
8
x



Early stopping could be myopic.

OVERFITTING

- **Overfitting:** high accuracy on training data, low accuracy on testing data

- Our learning algorithm performs a heuristic search
  - Overly complex functions are learned (e.g., those with as many parameters as data points)

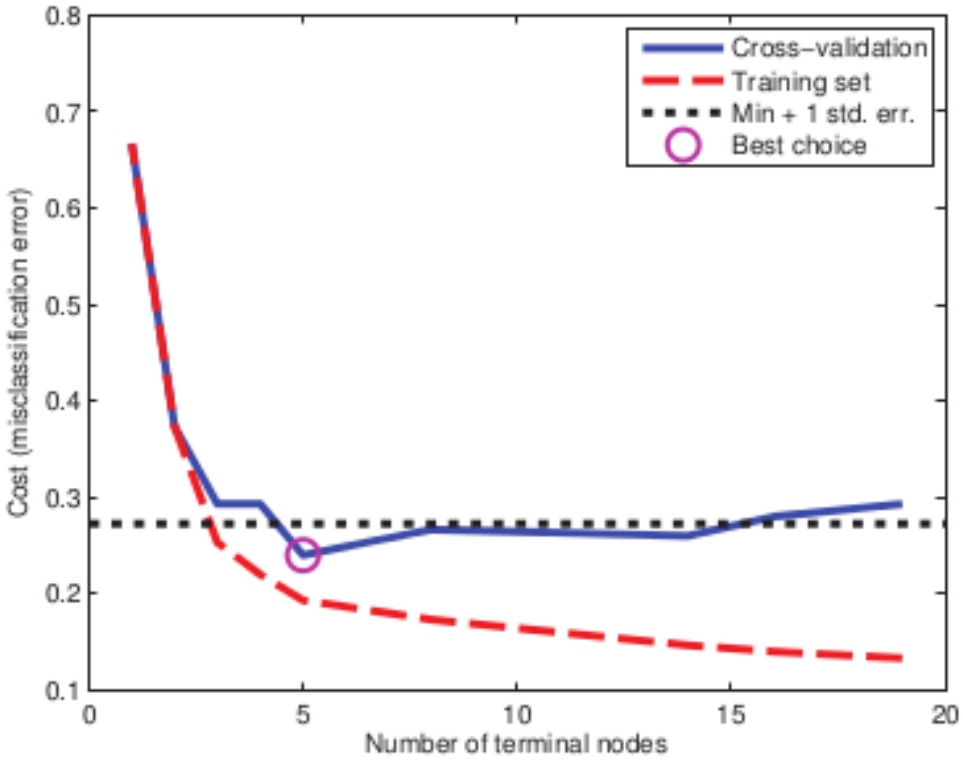
- How do we decide?

OCCAM'S RAZOR

- "Among competing hypotheses, the one that makes the fewest assumptions should be selected."

- Decision trees: the shorter the tree, the better

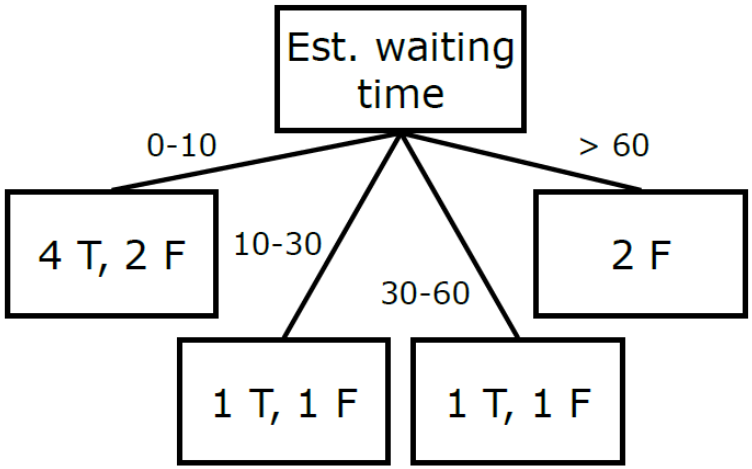
William of Ockham (1287-1347)


Cross−validation
Training set
Min + 1 std. err.
Best choice
Cost (misclassification error)
Number of terminal nodes
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20

BALANCING TRAINING SET PERFORMANCE AND MODEL COMPLEXITY

- Choose good splits early!
- Stop splitting when statistically insignificant
- Grow full tree and then prune afterwards
- Modify score function

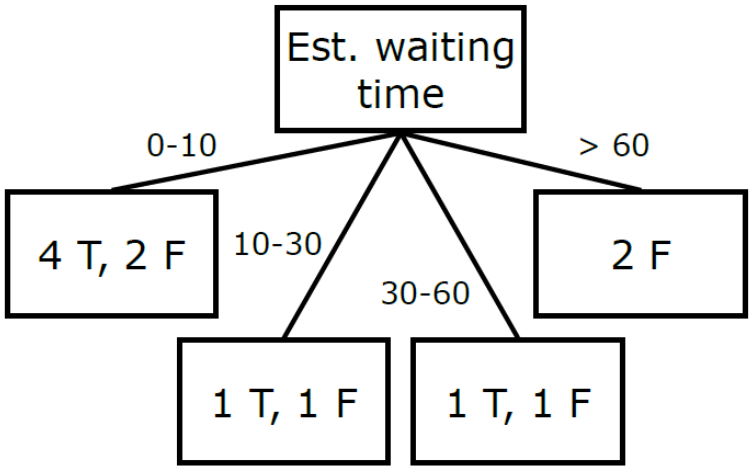
INFORMATION GAIN EXAMPLE



Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

$$\mathrm{IG}(A, D) = H(D) - \sum_{D_i \in \boldsymbol{D}} p(D_i) H(D_i)$$

INFORMATION GAIN EXAMPLE



Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

$$\mathrm{IG}(A, D) = H(D) - \sum_{D_i \in \boldsymbol{D}} p(D_i) H(D_i)$$

$$Remainder(Wait) = \frac{6}{12}\left[-\left(\tfrac{4}{6}\right)\log_2\left(\tfrac{4}{6}\right)-\left(\tfrac{2}{6}\right)\log_2\left(\tfrac{2}{6}\right)\right]$$

$$+\frac{2}{12}\left[-\left(\tfrac{1}{2}\right)\log_2\left(\tfrac{1}{2}\right)-\left(\tfrac{1}{2}\right)\log_2\left(\tfrac{1}{2}\right)\right]$$

$$+\frac{2}{12}\left[-\left(\tfrac{1}{2}\right)\log_2\left(\tfrac{1}{2}\right)-\left(\tfrac{1}{2}\right)\log_2\left(\tfrac{1}{2}\right)\right]+\frac{2}{12}\left[-\left(\tfrac{0}{2}\right)\log_2\left(\tfrac{0}{2}\right)-\left(\tfrac{2}{2}\right)\log_2\left(\tfrac{2}{2}\right)\right]=0.7925$$

Gain(Wait) = 1 – 0.7925 = 0.2075

DECISION TREE PLUSES AND MINUSES

SUMMARY

- Decision trees are very interpretable
- (Relatively) easy to implement
- Finding optimal decision tree is impractical
- In practice, less accurate than many other methods
- Often unstable
  - Small changes in training data lead to very different decision boundaries

ANNOUNCEMENTS

- Should have started on HW1
- Quiz on required Math and Probability background next time
  - Trying to book an additional room for it – will send a message on Piazza
- Asking questions on Piazza
  - Public (only way to get credit for participation)
  - Anonymous questions – if you prefer others not to know who asked
  - Private questions – only for private matters!
- Registration questions – posted an answer on Piazza

ACKNOWLEDGEMENTS

- The materials have been adapted from slides by Brian Ziebart and Marine Carpuat