Decision Trees: Representation

Machine Learning

THE UNIVERSITY OF UTAH

Some slides from Tom Mitchell, Dan Roth and others

Key issues in machine learning

- Modeling

  How to formulate your problem as a machine learning problem? How to represent data? Which algorithms to use? What learning protocols?

- Representation

  Good hypothesis spaces and good features

- Algorithms
  - What is a good learning algorithm?
  - What is success?
  - Generalization vs overfitting
  - The computational question: How long will learning take?

Coming up… (the rest of the semester)

Different hypothesis spaces and learning algorithms

- Decision trees and the ID3 algorithm
- Linear classifiers
  - Perceptron
  - SVM
  - Logistic regression
- Combining multiple classifiers
  - Boosting, bagging
- Non-linear classifiers
- Nearest neighbors

Coming up… (the rest of the semester)

Different hypothesis spaces and learning algorithms

– Decision trees and the ID3 algorithm
– Linea
  • Pe
  • SV
  • Lo
– Com
  • Bo
– Non-
– Near

Important issues to consider

1. What do these hypotheses represent?
2. Implicit assumptions and tradeoffs
3. Generalization?
4. How do we learn?

This lecture: Learning Decision Trees

1. **Representation**: What are decision trees?

2. **Algorithm**: Learning decision trees
   – The ID3 algorithm: A greedy heuristic

3. Some extensions

This lecture: Learning Decision Trees

1. **Representation**: What are decision trees?

2. **Algorithm**: Learning decision trees
   - The ID3 algorithm: A greedy heuristic

3. Some extensions

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Label
Claire Cardie	–
Peter Bartlett	+
Eric Baum	+
Haym Hirsh	–
Leslie Pack Kaelbling	+
Yoav Freund	–

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire Cardie	No	l	Yes	Yes	–
Peter Bartlett	No	e	No	No	+
Eric Baum	No	r	No	No	+
Haym Hirsh	No	a	No	Yes	–
Leslie Pack Kaelbling	No	e	Yes	No	+
Yoav Freund	No	o	No	No	–

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire					
Peter					
Eric E					
Haym Hirsh	No	a	No	Yes	−
Leslie Pack Kaelbling	No	e	Yes	No	+
Yoav Freund	No	o	No	No	−

With these four attributes, how many unique rows are possible?

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire					
Peter					
Eric E					
Haym Hirsh	No	a	No	Yes	−
Leslie Pack Kaelbling	No	e	Yes	No	+
Yoav Freund	No	o	No	No	−

With these four attributes, how many unique rows are possible?

$$2 \times 26 \times 2 \times 2 = 208$$

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire					
Peter					
Eric E					
Haym Hirsh	No	a	No	Yes	–
Leslie Pack Kaelbling	No	e	Yes	No	+
Yoav Freund	No	o	No	No	–

With these four attributes, how many unique rows are possible?

$$2 \times 26 \times 2 \times 2 = 208$$

If there are 100 attributes, all binary, how many unique rows are possible?

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire					
Peter					
Eric E					
Haym Hirsh	No	a	No	Yes	–
Leslie Pack Kaelbling	No	e	Yes	No	+
Yoav Freund	No	o	No	No	–

With these four attributes, how many unique rows are possible?

$2 \times 26 \times 2 \times 2 \ = \ 208$

If there are 100 attributes, all binary, how many unique rows are possible?

(100 times) $2 \times 2 \times 2 \times \cdots \times 2 = 2^{100}$

Representing data

Data can be represented as a big table, with columns denoting different attributes

Name	Name has punctuation?	Second character of first name	Length of first name>5?	Same first letter in two names?	Label
Claire					
Peter					
Eric E					
Haym Hirsh	No	a	No	Yes	–

With these four attributes, how many unique rows are possible?

$2 \times 26 \times 2 \times 2 = 208$

If there are 100 attributes, all binary, how many unique rows are possible?

(100 times) $2 \times 2 \times 2 \times \cdots \times 2 = 2^{100}$

If we wanted to store all possible rows, this number is too large.

We need to figure out how to represent data in a better, more efficient way

What are decision trees?

A hierarchical data structure that represents data using a divide-and-conquer strategy

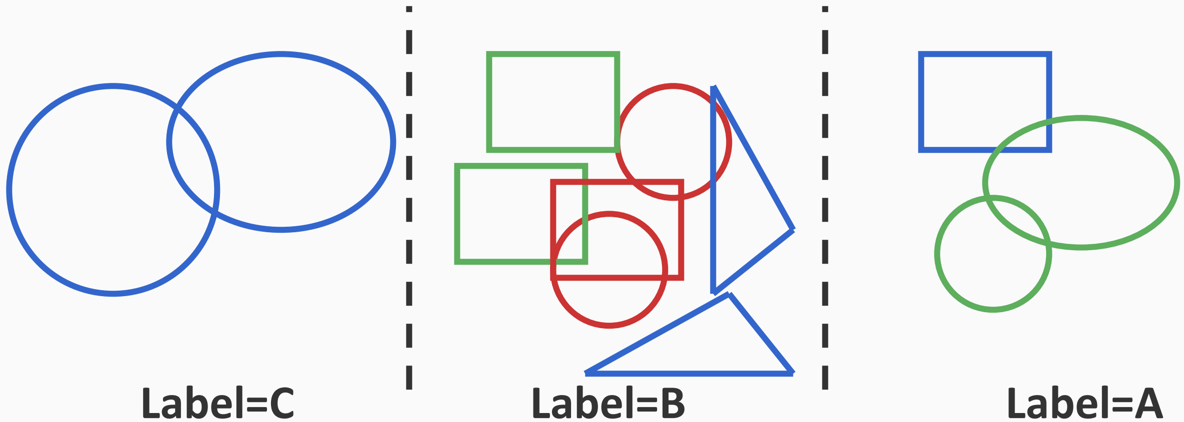
Can be used as hypothesis class for non-parametric classification or regression

General idea: Given a collection of examples, learn a decision tree that represents it

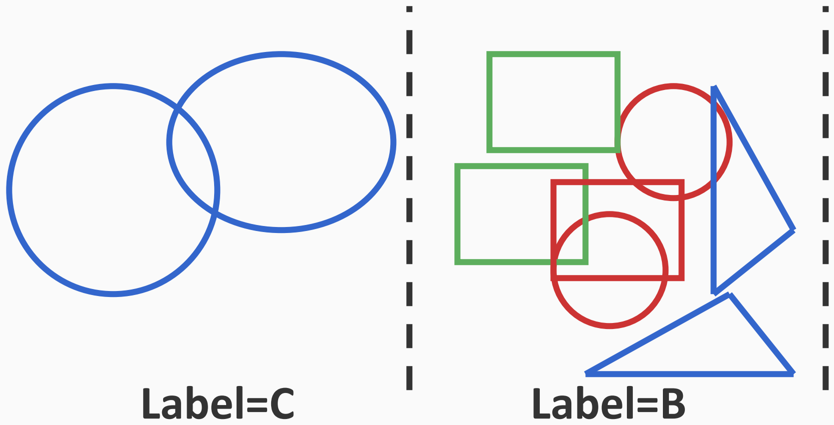
What are decision trees?

- Decision trees are a family of classifiers for instances that are represented by collections of attributes (i.e. features)
- Nodes are tests for feature values
- There is one branch for every value that the feature can take
- Leaves of the tree specify the class labels

Let’s build a decision tree for classifying shapes



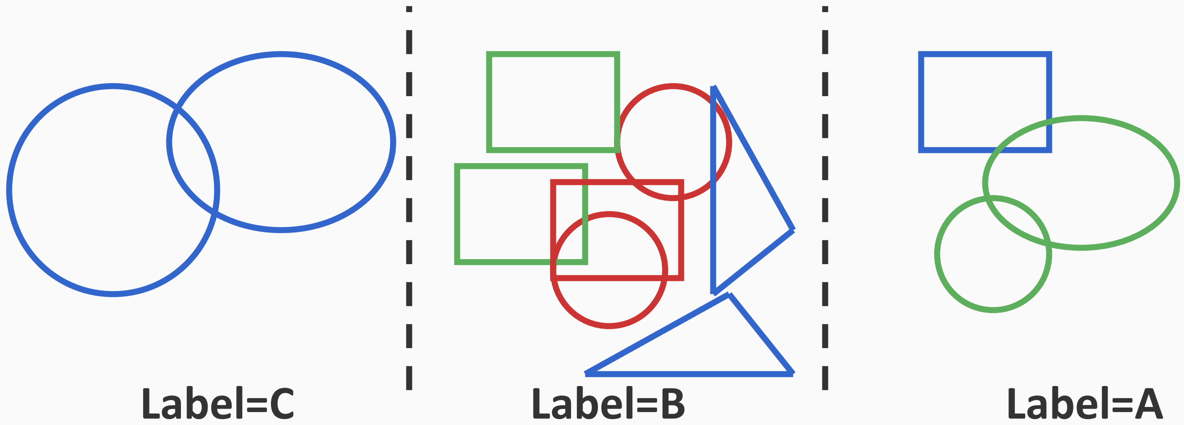
Let’s build a decision tree for classifying shapes



Before building a decision tree:

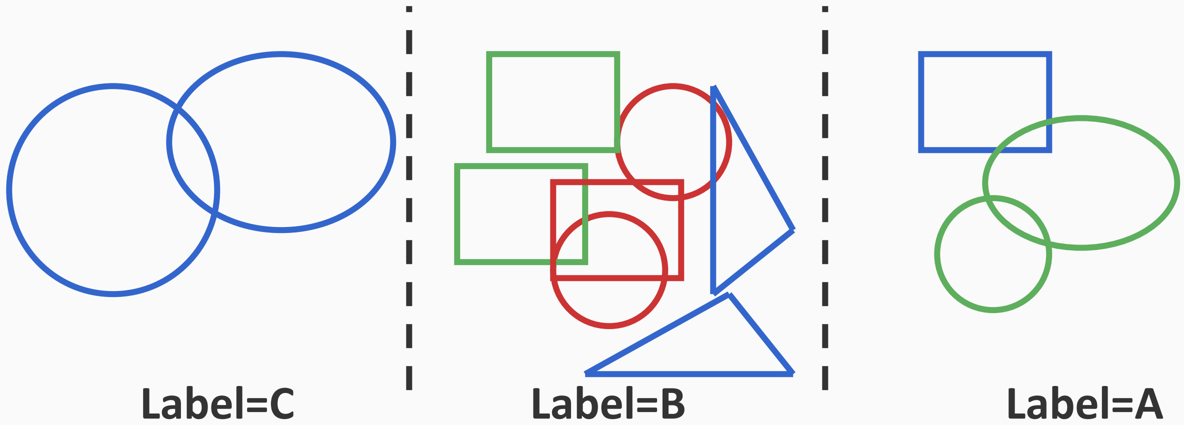
What is the label for a red triangle? And why?

Let’s build a decision tree for classifying shapes



What are some attributes of the examples?

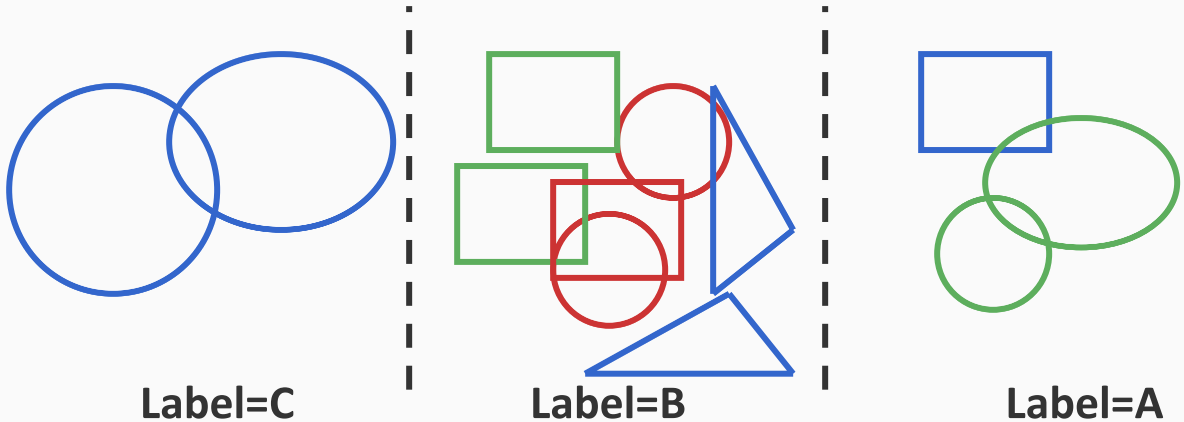
Let’s build a decision tree for classifying shapes



What are some attributes of the examples?

Color, Shape

Let’s build a decision tree for classifying shapes

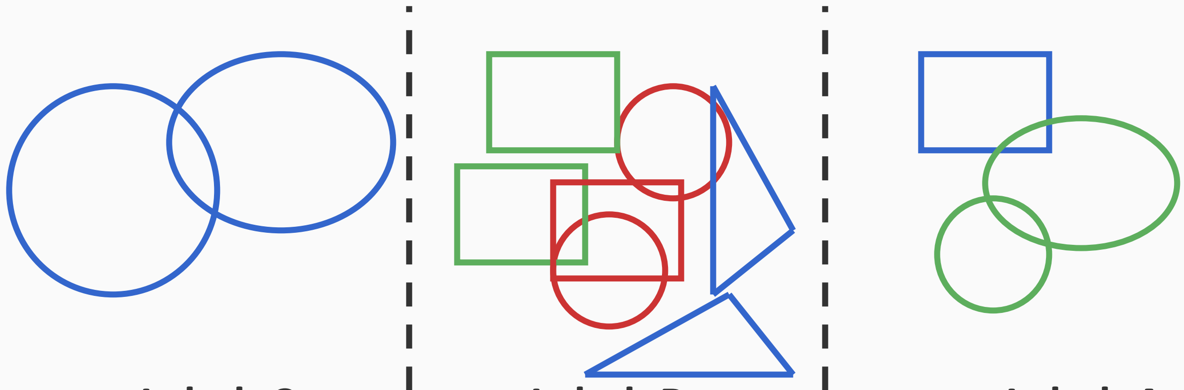


What are some attributes of the examples?

Color, Shape

$\boxed{\textbf{Color?}}$

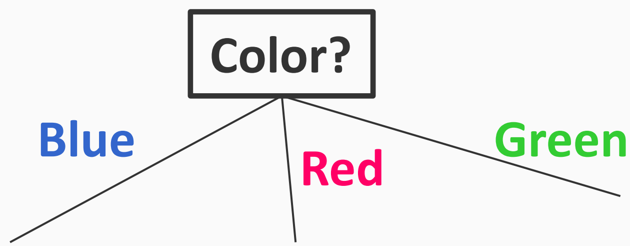
Let’s build a decision tree for classifying shapes



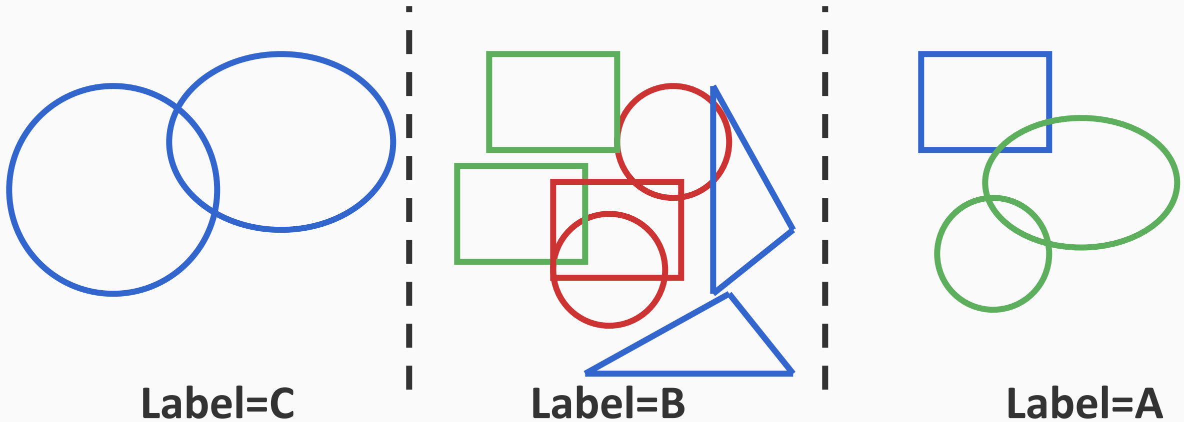
Label=C Label=B Label=A

What are some attributes of the examples?

Color, Shape

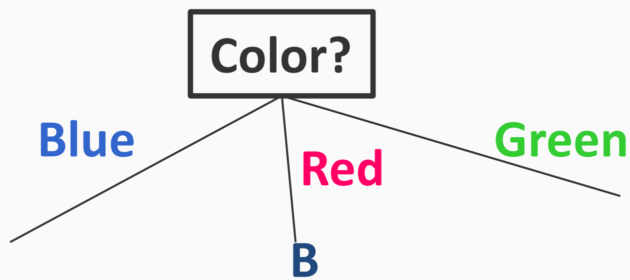


Let’s build a decision tree for classifying shapes

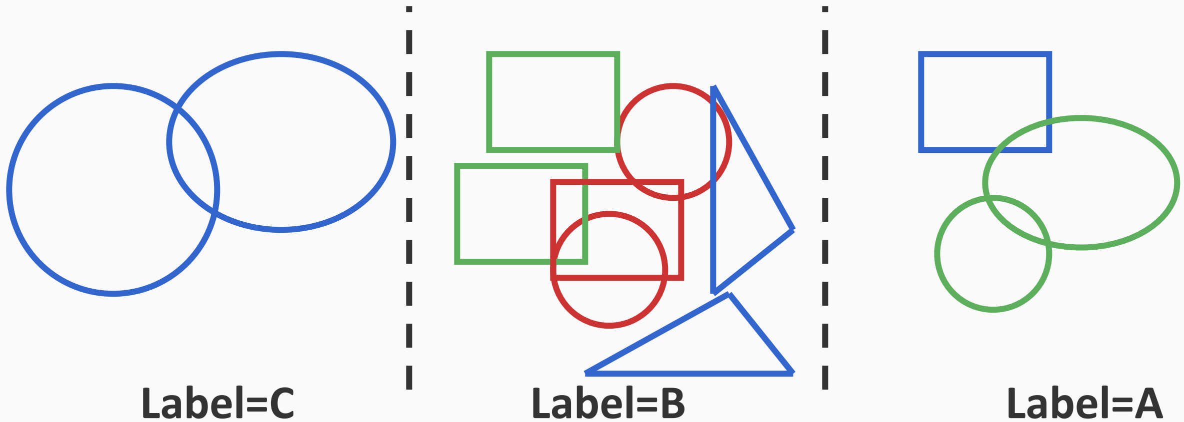


What are some attributes of the examples?

Color, Shape

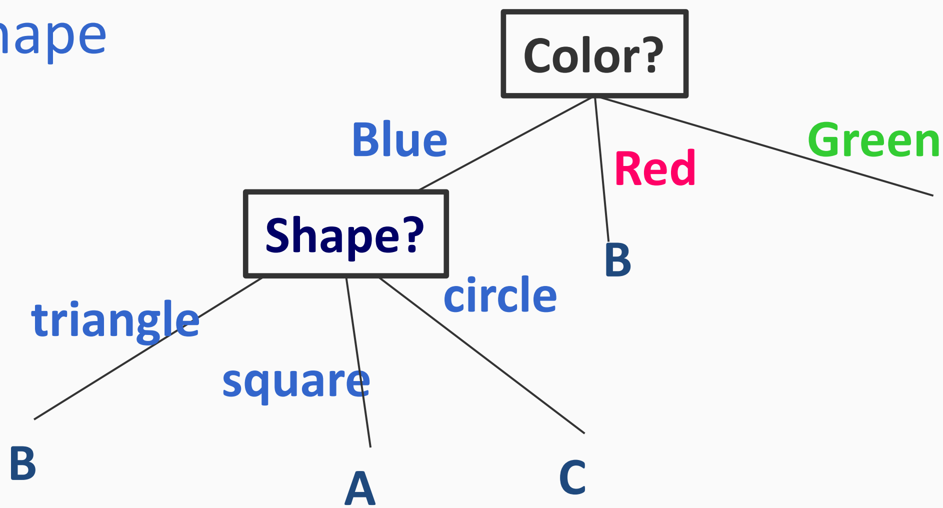


Let’s build a decision tree for classifying shapes

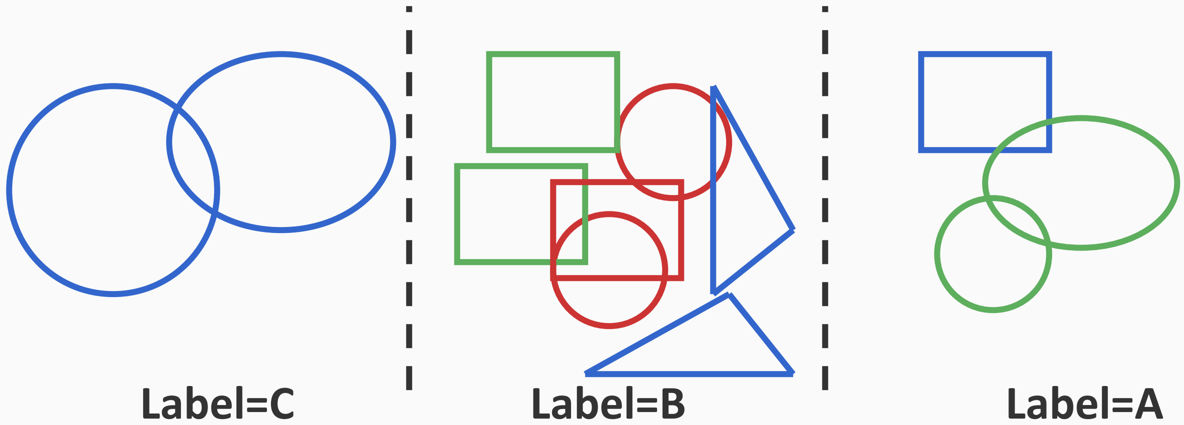


What are some attributes of the examples?

Color, Shape

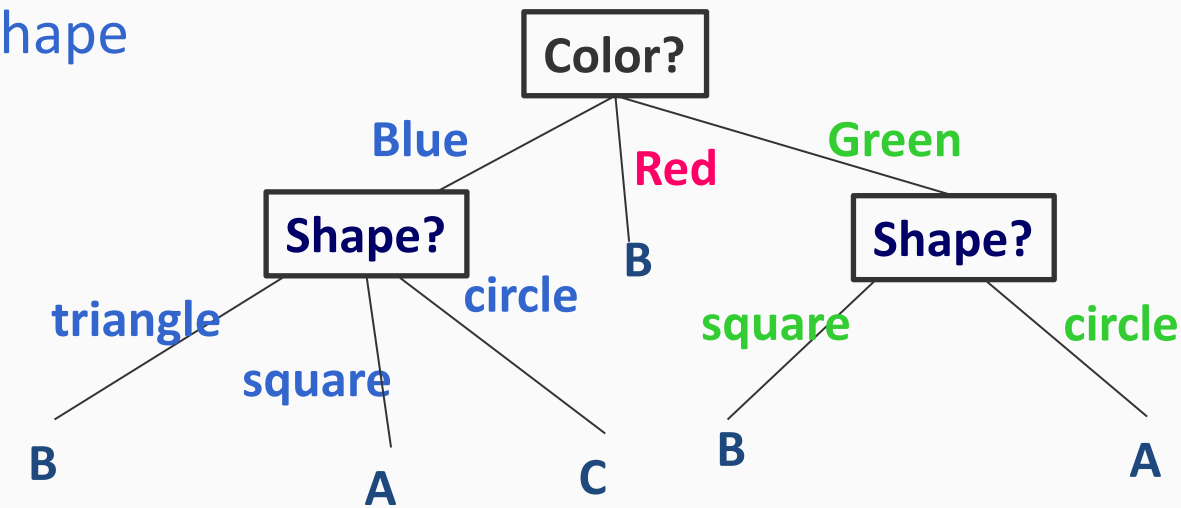


Let’s build a decision tree for classifying shapes



What are some attributes of the examples?

Color, Shape

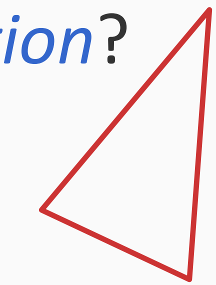


Let’s build a decision tree for classifying shapes

1. How do we *learn* a decision tree?

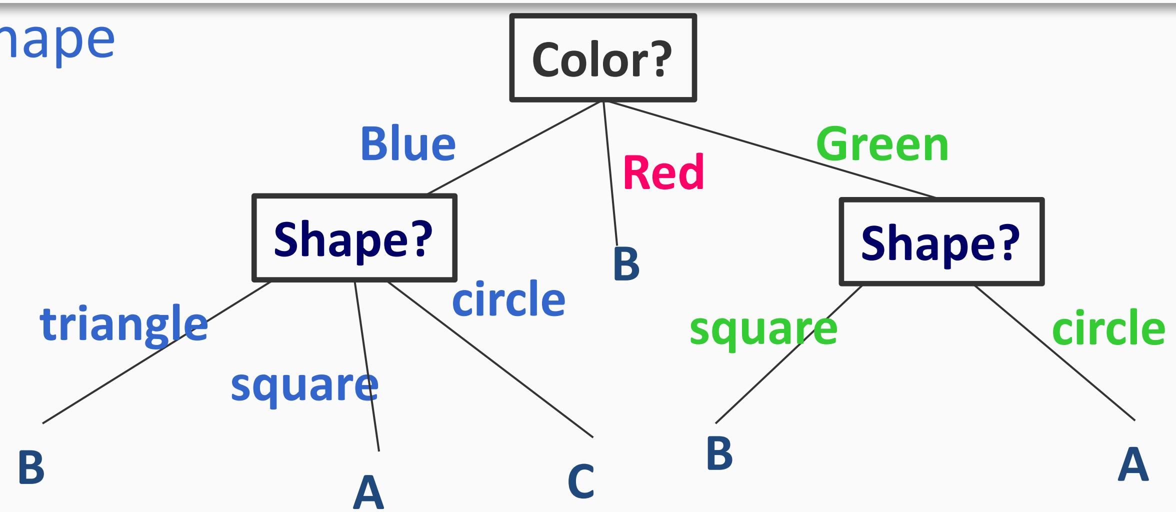
   Coming up soon…

2. How to use a decision tree for *prediction*?
   - What is the label for a **red** triangle?
     - Just follow a path from the root to a leaf



Wh

Color, Shape

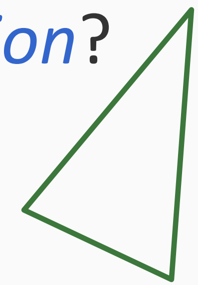


Let's build a decision tree for classifying shapes

1. How do we *learn* a decision tree?

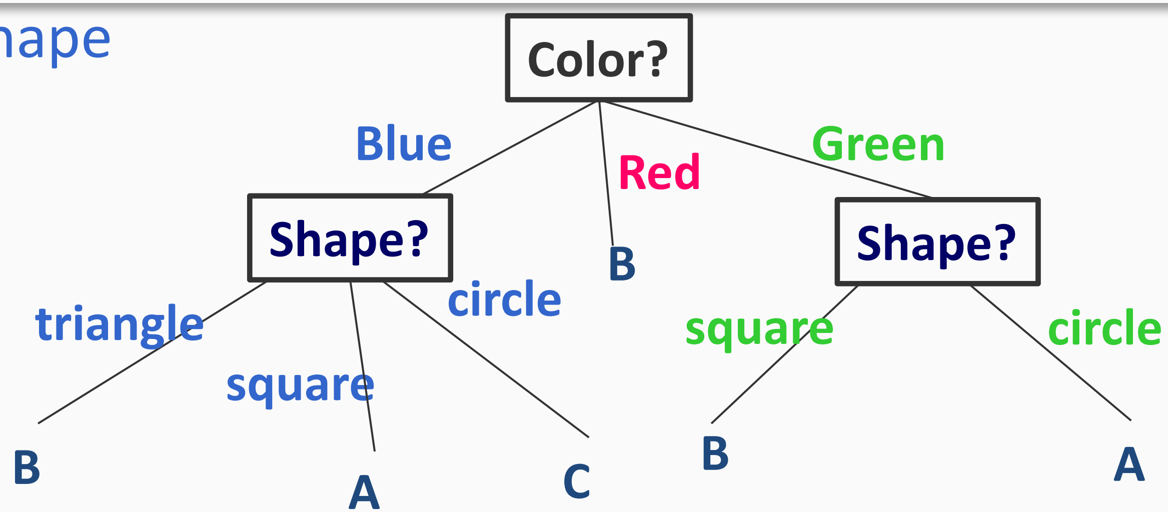
   Coming up soon…

2. How to use a decision tree for *prediction*?
   - What is the label for a **red** triangle?
     - Just follow a path from the root to a leaf
   - What about a **green** triangle?



Wh

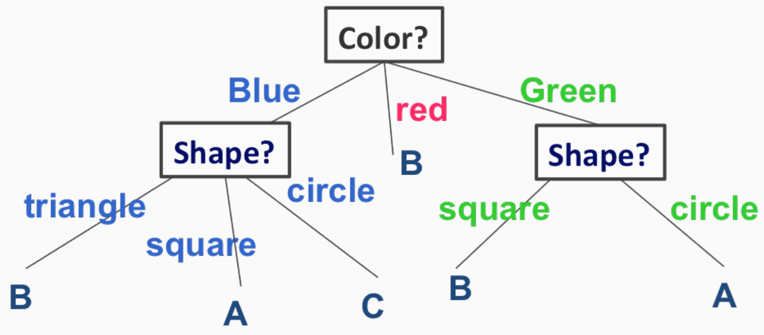
Color, Shape



Expressivity of Decision trees

What Boolean functions can decision trees represent?

– *Any* Boolean function



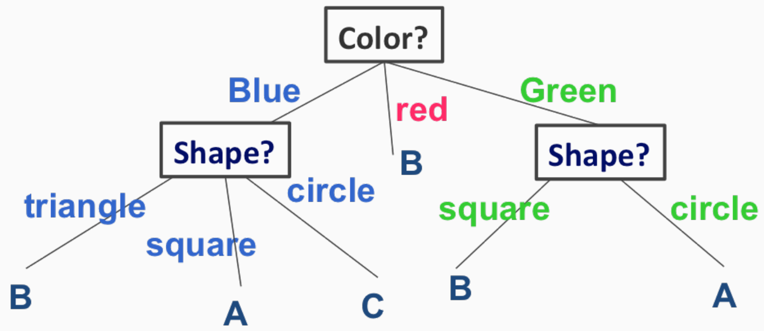
Every path from the tree to a root is a rule

The full tree is equivalent to the conjunction of all the rules

Expressivity of Decision trees

What Boolean functions can decision trees represent?

- *Any* Boolean function



Every path from the tree to a root is a rule

The full tree is equivalent to the conjunction of all the rules

(Color=blue AND Shape=triangle) ⇒ Label=B) AND

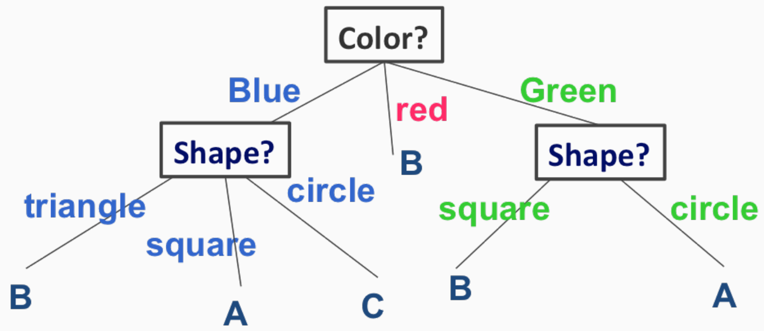
(Color=blue AND Shape=square) ⇒ Label=A) AND

(Color=blue AND Shape=circle) ⇒ Label=C) AND….

Expressivity of Decision trees

What Boolean functions can decision trees represent?

– *Any* Boolean function



Every path from the tree to a root is a rule

The full tree is equivalent to the conjunction of all the rules

(Color=blue AND Shape=triangle) ⇒ Label=B) AND

(Color=blue AND Shape=square) ⇒ Label=A) AND

(Color=blue AND Shape=circle) ⇒ Label=C) AND….

Any Boolean function can be represented as a decision tree.

Decision Trees

- Outputs are discrete categories
- But real valued outputs are also possible (regression trees)
- Well studied methods for handling noisy data (noise in the label or in the features) and for handling missing attributes
  - Pruning trees helps with noise
  - More on this later…

Numeric attributes and decision boundaries

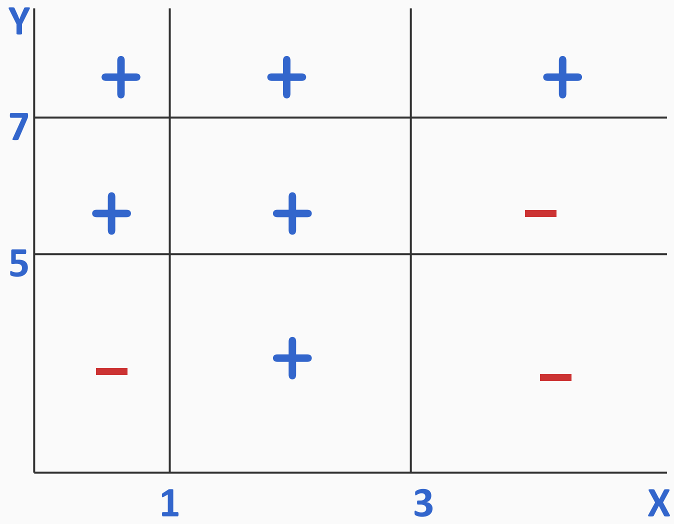
- We have seen instances represented as attribute-value pairs (color=blue, second letter=e, etc.)
  - Values have been categorical
- How do we deal with numeric feature values? (eg length = ?)

Numeric attributes and decision boundaries

- We have seen instances represented as attribute-value pairs (color=blue, second letter=e, etc.)
  - Values have been categorical
- How do we deal with numeric feature values? (eg length = ?)
  - Discretize them or use thresholds on the numeric values

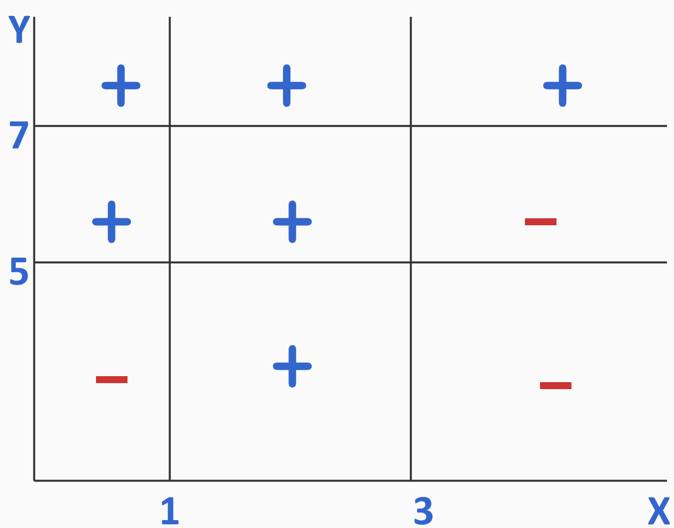
Numeric attributes and decision boundaries

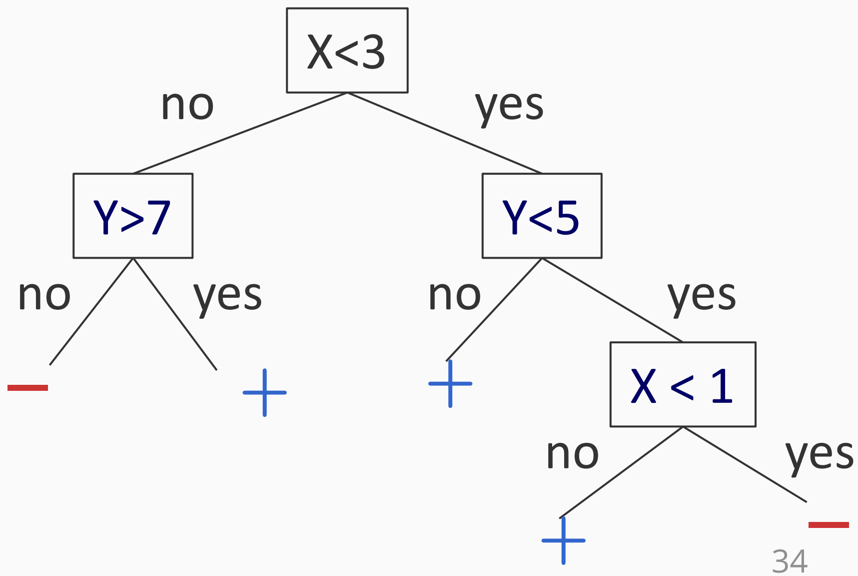
- We have seen instances represented as attribute-value pairs (color=blue, second letter=e, etc.)
  - Values have been categorical
- How do we deal with numeric feature values? (eg length = ?)
  - Discretize them or use thresholds on the numeric values
  - This example divides the feature space into axis parallel rectangles



Numeric attributes and decision boundaries

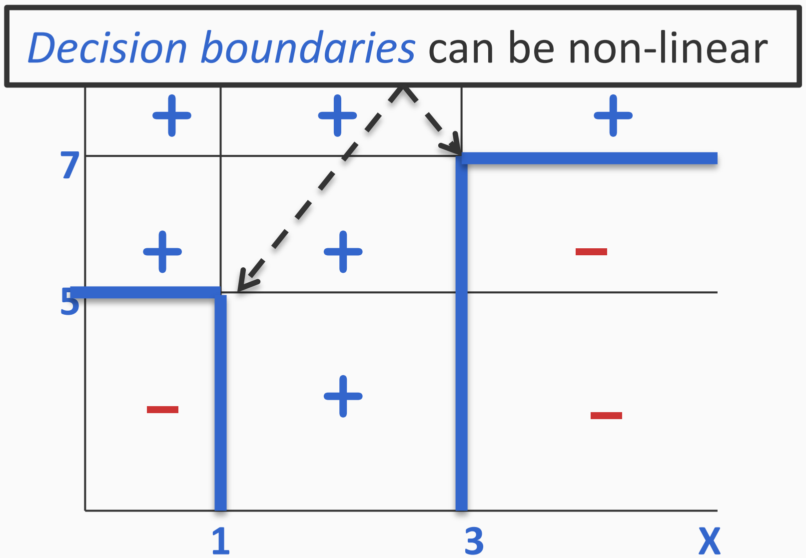
- We have seen instances represented as attribute-value pairs (color=blue, second letter=e, etc.)
  - Values have been categorical
- How do we deal with numeric feature values? (eg length = ?)
  - Discretize them or use thresholds on the numeric values
  - This example divides the feature space into axis parallel rectangles

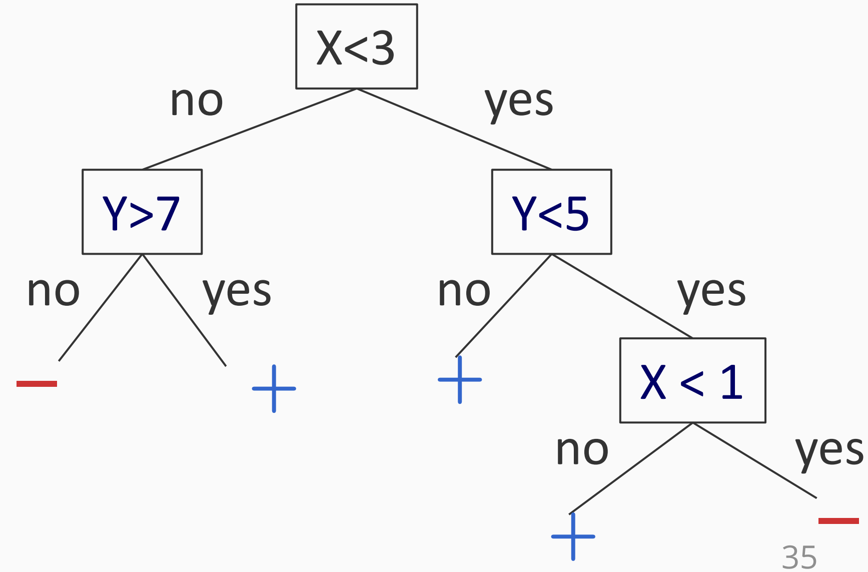




Numeric attributes and decision boundaries

- We have seen instances represented as attribute-value pairs (color=blue, second letter=e, etc.)
  - Values have been categorical
- How do we deal with numeric feature values? (eg length = ?)
  - Discretize them or use thresholds on the numeric values
  - This example divides the feature space into axis parallel rectangles





Summary: Decision trees

- Decision trees can represent any Boolean function
- A way to represent lot of data
- A natural representation (think 20 questions)
- Predicting with a decision tree is easy

- Clearly, given a dataset, there are many decision trees that can represent it. [Exercise: Why?]

Summary: Decision trees

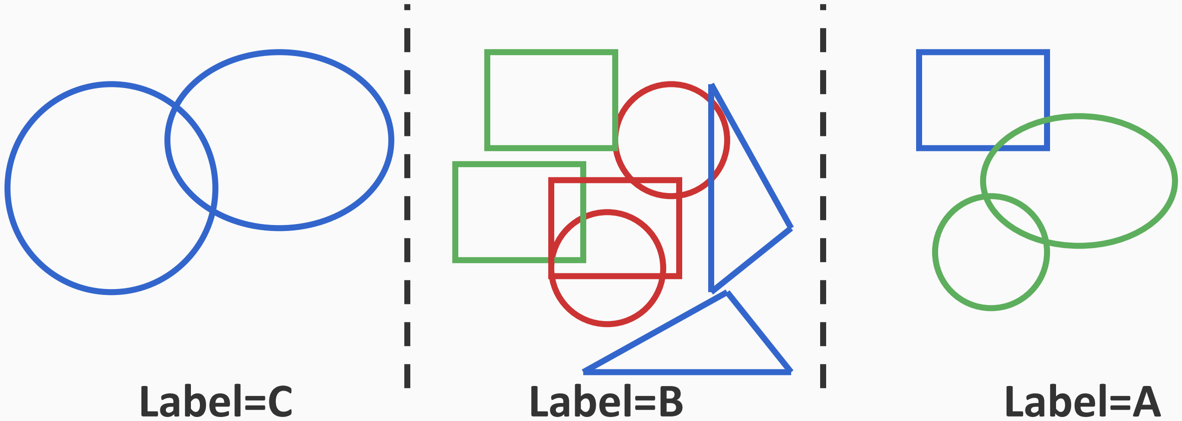
- Decision trees can represent any Boolean function
- A way to represent lot of data
- A natural representation (think 20 questions)
- Predicting with a decision tree is easy

- Clearly, given a dataset, there are many decision trees that can represent it. [Exercise: Why?]

- Learning a good representation from data is the next question

Exercises

1. Write down the decision tree for the shapes data if the root node was *Shape* instead of *Color.*
2. Will the two trees make the same predictions for unseen shapes/color combinations?



3. Show that multiple structurally different decision trees can represent the same Boolean function of two or more variables.

   (think about what it means for two trees to be structurally different)