

# Decision Trees & Supervised Learning

by Adam Hawkins for CS157B

# Supervised Learning

Supervised learning is also called classification or inductive learning. This is similar to human learning from past experiences to gain new knowledge in order to improve our ability to perform real world tasks. However, since machines do not have “experiences,” machines learn from data that represent past experiences.

# An Experience

<b>Age</b>	<b>Has_job</b>	<b>Own_House</b>	<b>Credit</b>	<b>Class</b>
young	FALSE	FALSE	fair	no
young	TRUE	FALSE	good	yes
middle	FALSE	FALSE	fair	no
middle	TRUE	TRUE	good	yes
old	FALSE	TRUE	excellent	yes
old	FALSE	TRUE	good	yes
old	FALSE	FALSE	fair	no

# What's the Goal?

- Have the computer determine some function that knows the result given some input data.
- This function is called the classification model.
- The data used to determine the function is called training data.
- We can test our function using test data. This will tell us how accurate our classification model is.
- One way to generate the classification model is a Decision Tree.

# What is a decision tree?

- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including a chance event outcomes, resource costs, and utility.

Wait, What?

I've got 20 questions

# Use a Decision Tree



Step Through the Tree to Solve the Problem

# Our Experience

- Using the table from the previous slides we want to determine some rules like this:
- $\text{Own\_house} = \text{true}, \Rightarrow \text{Answer:Yes}$
- $\text{Own\_house} = \text{false}, \text{Has\_Job} = \text{true}, \Rightarrow \text{Answer:Yes}$
- $\text{Own\_house} = \text{false}, \text{Has\_Job} = \text{false}, \Rightarrow \text{Answer: No}$

# Generating The Rules

Learning a tree is typically done using a divide-and-conquer strategy that recursively partitions the data to produce the tree. At the beginning, all the examples are at the root. As the tree grows, the examples are sub-divided recursively.

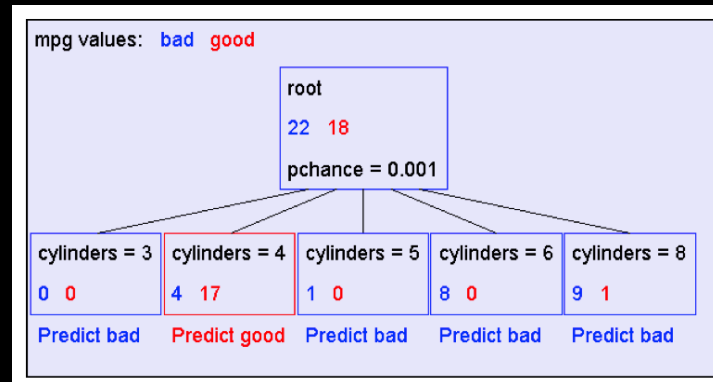
# BuildTree(DataSet, Out)

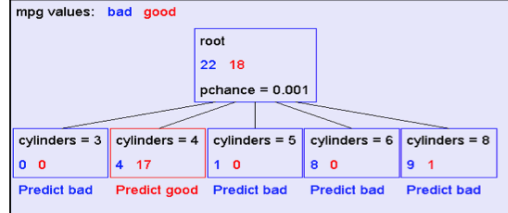
- If all the output values are the same in DataSet, return a leaf node that says “predict this unique output”
- If all the input values are the same, return a leaf node that says “predict the majority output”
- Else find attribute  $X$  with highest Info Gain
- Suppose  $X$  has  $N$  distinct values:
  - Create and return a non-leaf node with  $X$  children
    - the  $i$ th child should be built using: BuildTree(DS, Output).  
Where DS consists of all those records in DataSet for which  $X = i$ th distinct value of  $X$

# An Example

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

# Building a Tree

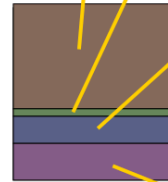




Take the Original Dataset..



And partition it according to the value of the attribute we split on

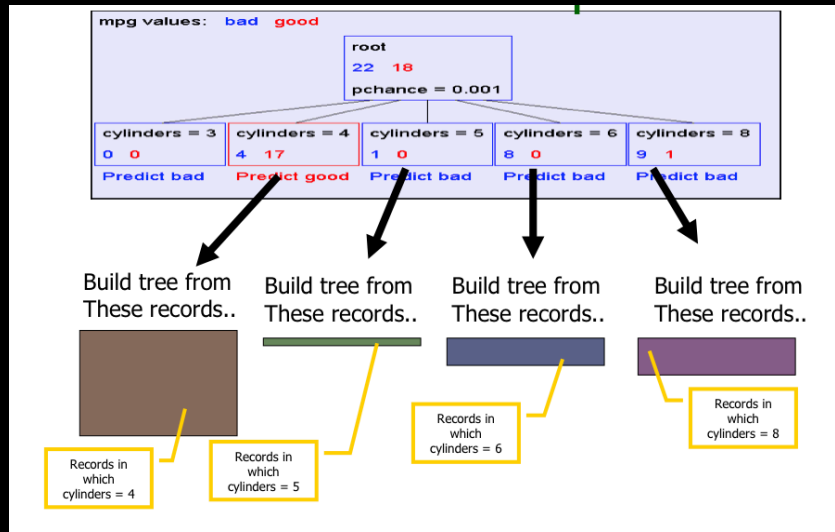


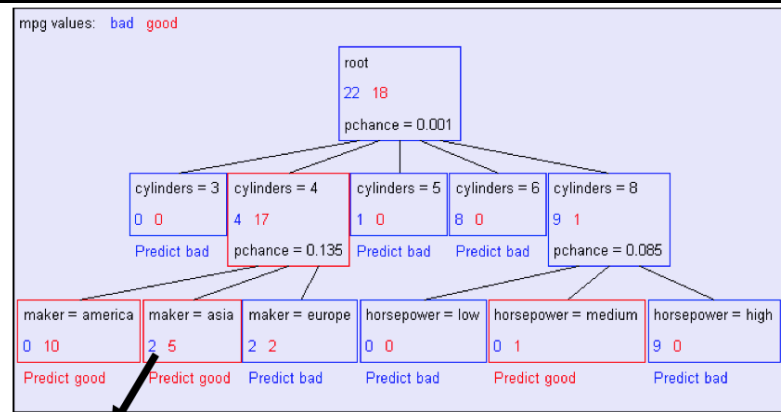
Records in which cylinders = 4

Records in which cylinders = 5

Records in which cylinders = 6

Records in which cylinders = 8

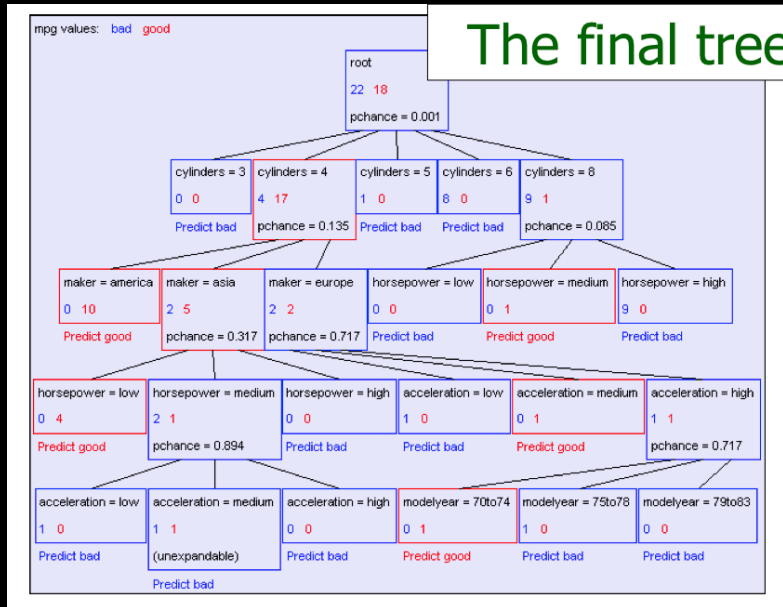




Recursively build a tree from the seven records in which there are four cylinders and the maker was based in Asia

(Similar recursion in the other cases)

# The final tree



# What's Next?

- Run through a test data set and see how accurate your tree is

# Why are we doing this?

- You may be thinking, wow that's great we can make a tree to model data we already know the answer to. But, now we can use our tree to predict FUTURE data.

# TL;DR

- Supervised learning is trying to teach a machine some thing by using existing data with known answers.
- Decision Trees are a way to classify existing data in a way that we can predict future results.
- Construct DT's from datasets using recursion.
- Test your DT against some test Data
- .....
- Profit!?!?