SUPERVISED LEARNING + DECISION TREES

Progress Report

- We've finished Part I: Problem Solving
- We've finished Part II: Reasoning with uncertainty!
- Part III: (Machine) Learning
  - Supervised Learning
  - Unsupervised Learning
  - (Reinforcement Learning)
- Overlaps quite a bit with Part II

Today

- Reading
  - We're skipping to AIMA Chapter 18!
  - AIMA 18.1-18.4
- Goals
  - What is machine learning?
  - What is supervised learning?
  - Decision trees

Machine Learning

- The goal of machine learning is to learn from data
  - We might use machine learning to
    - learn the probabilities for a Bayesian network
    - learn the topology of a Bayesian network
- Three types of learning
  - Supervised learning – learning with labels
  - Unsupervised learning – learning without labels
  - Reinforcement learning – learning with rewards

Supervised learning terminology

$$\mathbf{x}_i \longrightarrow \boxed{f} \longrightarrow y_i$$

- Training set

$$D = \{(\mathbf{x}_i, y_i) \mid i = 1, \dots, N\} \quad \text{where} \quad f(\mathbf{x}_i) = y_i$$

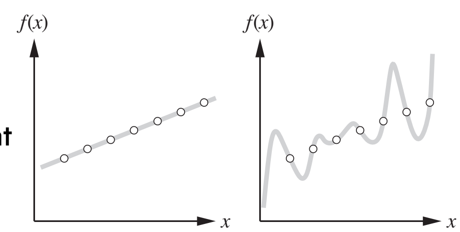
- Hypothesis class

$$h \in \mathcal{H}$$

- Given training set, we want to find the hypothesis in the hypothesis class that "best approximates" f

Supervised learning terminology

- Example: Curve fitting
  - x is the x-coordinate
  - y is the y-coordinate
  - Both hypotheses are consistent
  - Which is better?



(a) (b)

- Ockham's Razor
  - Prefer the simplest consistent hypothesis
- Test set
  - Evaluate performance of each hypothesis on a new (unseen) set of examples

Supervised Learning terminology

- Regression
  - y is a real-valued number
  - e.g. price of a commodity, pollution levels, brain activity
- Classification
  - y is a discrete (categorical) value
  - e.g. spam or not spam, 5-star ratings
- Structured prediction
  - y is a structured object
  - e.g. given sentence predict parse tree, given words in a sentence predict POS tags

Supervised Learning

- Learning with labels
  - Spam
  - Digit recognition
  - Rainfall levels in India
  - Pollution index
  - Stock returns
  - User's ratings of movies
  - Genre classification
  - Sentiment analysis
  - Document classification
  - Image recognition
  - Part-of-speech
  - Storm trajectories

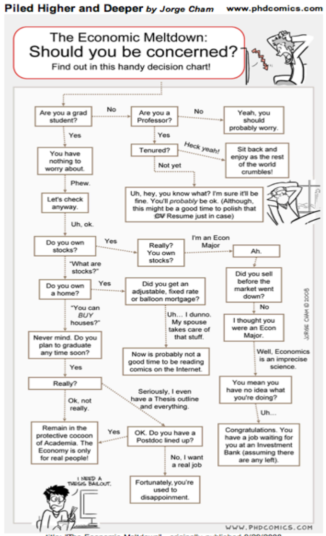
Common Supervised Learning Algorithms

- Graphical models
  - Naïve Bayes classifiers
  - Bayesian networks
- Decision trees
  - Random forests (many decision trees)
- Neural Networks
  - Perceptrons
  - Artificial neural networks
  - Deep belief nets
- Max margin classifiers
  - Support vector machines
- Regression analysis
  - Logistic regression
  - Linear regression

A procedure for taking a set of labeled examples (i.e. the training set), and constructing a hypothesis h that has the best performance on the training set.

Decision trees

Piled Higher and Deeper *by Jorge Cham* www.phdcomics.com

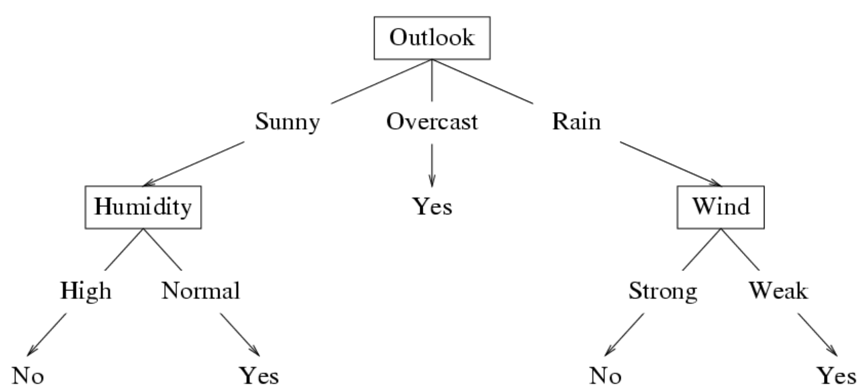


title: "The Economic Meltdown" - originally published 9/29/2008

Decision trees

	Day	Outlook	Temp.	Humidity	Wind	PlayTennis	
$\mathbf{x}_1 \longrightarrow$	D1	Sunny	Hot	High	Weak	No	$\longleftarrow y_1$
$\mathbf{x}_2 \longrightarrow$	D2	Sunny	Hot	High	Strong	No	$\longleftarrow y_2$
$\mathbf{x}_3 \longrightarrow$	D3	Overcast	Hot	High	Weak	Yes	$\longleftarrow y_3$
	D4	Rain	Mild	High	Weak	Yes	
	D5	Rain	Cool	Normal	Weak	Yes	
	D6	Rain	Cool	Normal	Strong	No	
	D7	Overcast	Cool	Normal	Strong	Yes	
	D8	Sunny	Mild	High	Weak	No	
	D9	Sunny	Cool	Normal	Weak	Yes	
	D10	Rain	Mild	Normal	Weak	Yes	
	D11	Sunny	Mild	Normal	Strong	Yes	
	D12	Overcast	Mild	High	Strong	Yes	
	D13	Overcast	Hot	Normal	Weak	Yes	
	D14	Rain	Mild	High	Strong	No	

Decision Trees

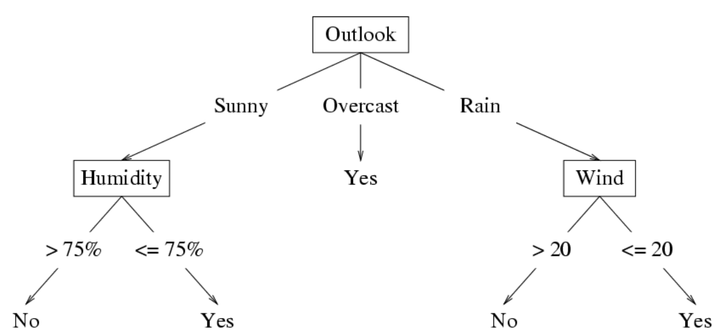


Decision Trees

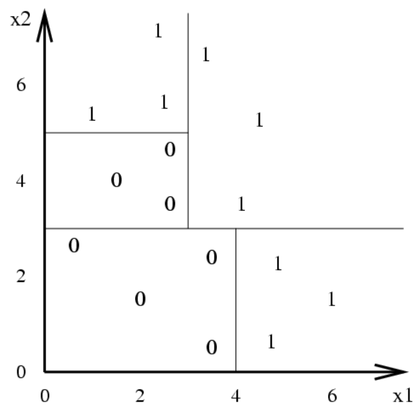
- Decision trees are best suited to problems where
  - Each attribute is discrete
  - The label y is discrete
  - The hypothesis can be expressed using conjunctions (AND) and disjunctions (OR)
  - The training data may contain errors
  - The training data may contain missing attribute values

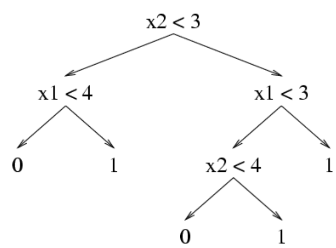
Decision Trees

- If the features are continuous, internal nodes may test the value of a feature against a threshold

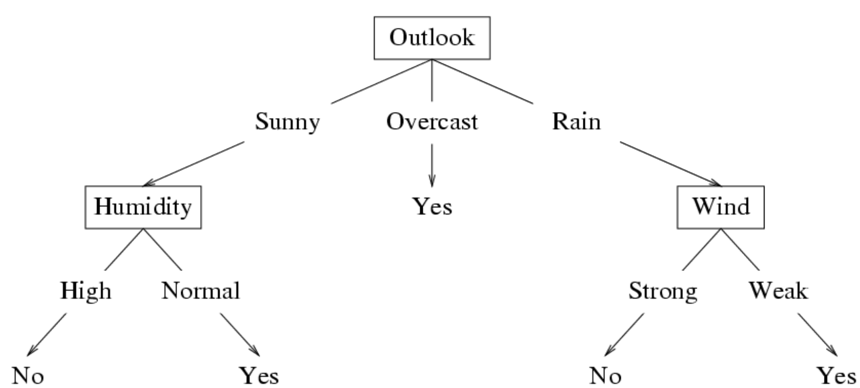


Decision Trees





Learning a Decision Tree



Learning a Decision Tree

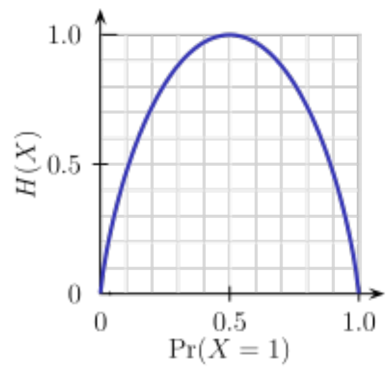
```
function DECISION-TREE-LEARNING (examples, attributes, parents) returns a tree
    if examples is empty return MAJORITY_VOTE(parents)
    else if all examples have same classification return classification
    else if attributes is empty return MAJORITY_VOTE(examples)
    else
        A    ←——   CHOOSE-BEST-ATTRIBUTE (examples)
        tree ←——   a new decision tree with root A
        for each value v_k of A
            S_k      ←——  examples with value v_k for attribute A
            subtree ←——  DECISION-TREE-LEARNING(S_k, attributes-A, examples)
            add branch to tree with label (A=v_k) and subtree
        return tree
```

Choosing the best attribute

- Splitting on a good attribute
  - After the split, the examples at each branch have the same classification
- Splitting on a bad attribute
  - After the split, the examples at each branch have the same proportion of positive and negative examples
- We will use entropy and information gain to formalize what we mean by *good* and *bad* attributes

Entropy

- **Entropy** measures the uncertainty of a random variable
  - How many bits are needed to efficiently encode the possible values (outcomes) of a random variable?
- Introduced by Shannon in 1948 paper
- Example: flipping a coin
  - A completely biased coin requires 0 bits of entropy
  - A fair coin requires 1 bit of entropy
  - How many bits are need to encode the outcome of flipping a fair coin twice?



Entropy and Information Gain

- Let A be a random variable with values v_k
- Each value v_k occurs with probability $p(v_k)$
- Then the entropy of A is defined as

$$H(A) = \sum_k p(v_k)\, log_2\left(\frac{1}{p(v_k)}\right)$$
$$= -\sum_k p(v_k)\, log_2\, p(v_k)$$

- (Apply this notion of entropy to choosing the best attribute)

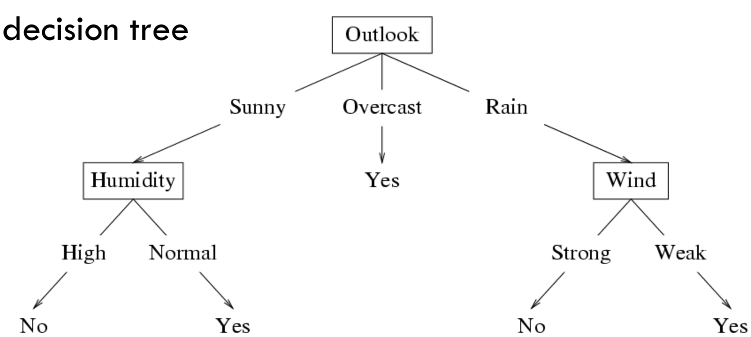
Entropy and Information Gain

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

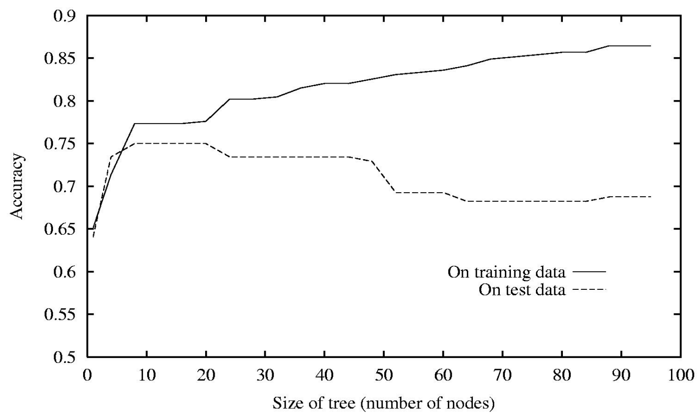
Day	Outlook	Temp.	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Decision Trees: additional considerations

- Overfitting can be caused by many factors
  - Noisy data, irrelevant attributes, spurious correlations, non-determinism
- Can cause additional nodes to be added to the decision tree



Decision Trees: additional considerations



Decision Trees: additional considerations

- Overfitting
  - Can post-process the learned decision tree and prune using significance testing at final nodes
  - Cross-validation using validity set
- Continuous or integer-valued attributes
  - Use ranges
- Continuous label y
  - Combination of splitting and linear regression