SUPERVISED LEARNING

Progress Report

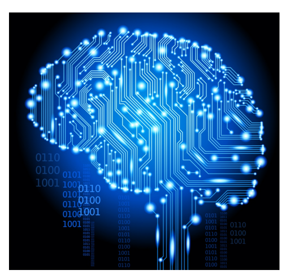
- We've finished Part I: Problem Solving
- We've finished Part II: Reasoning with uncertainty
- Part III: (Machine) Learning
  - Supervised Learning
  - Unsupervised Learning
- Overlaps quite a bit with Part II

Today

- Reading
  - We're skipping to AIMA Chapter 18!
  - AIMA 18.1-18.2, *skim* 20.2.2
- Goals
  - Intro. to Machine Learning
  - Supervised learning terminology
  - Naïve Bayes
  - (Decision Trees)

Machine Learning

- The term "machine learning" is a bit misleading
  - Pattern recognition
- We can use machine learning to
  - learn the probabilities for a BN
  - learn the topology of a BN
  - learn heuristic function for games



Subfields of Machine Learning

- Supervised learning
  - learning with labels
  - classification, regression, structured prediction
- Unsupervised learning
  - learning without labels
  - clustering, projection methods
- Reinforcement learning
  - learning with rewards
  - planning

Supervised Learning Terminology

- data set
- instance, input
- features
- label, output
- hypothesis
- hypothesis class
- realizable, consistent

Types of Supervised Learning Tasks

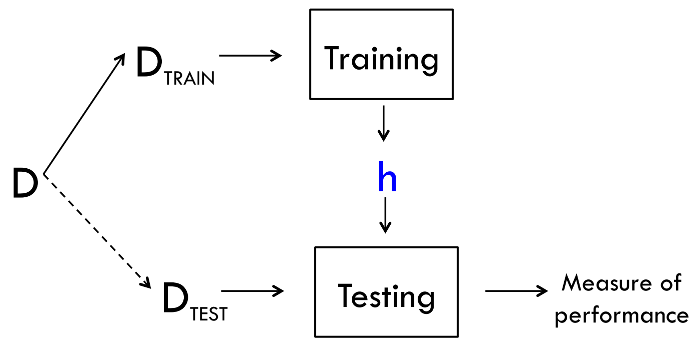
- Regression
  - y is a (vector of) real-valued number(s)
  - e.g. price of a commodity, pollution levels, brain activity
- Classification
  - y is a discrete (categorical) value
  - e.g. spam or not spam, 5-star ratings
- Structured prediction
  - y is a structured object
  - e.g. given sentence predict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

Types of Supervised Learning Tasks

- Supervised learning
  - 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

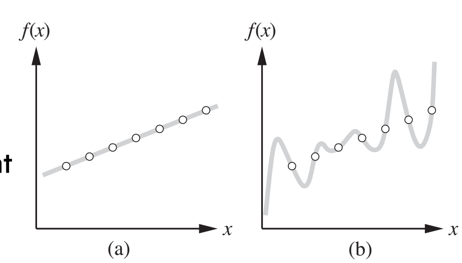
So what is learning?

- Learning is the process of finding (constructing, searching for) a hypothesis that performs well on the training data and generalizes well to unseen data (the test data)



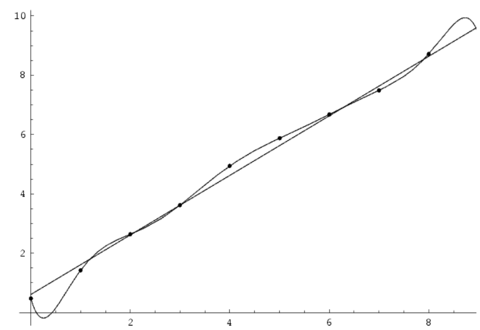
Ockham's Razor (inductive bias)

- Ockham's Razor
  - Prefer the simplest consistent hypothesis
- Example: Curve fitting
  - x is the x-coordinate
  - y is the y-coordinate
  - Both hypotheses are consistent
  - Which is better?



Overfitting (phenomenon)

- Overfitting
  - Learner fits itself to noise in the training data failing to generalize well
  - Causes: noisy data (too little data), overly complex models
- Example: Curve fitting
  - Which is better?



Common Supervised Learning Algorithms

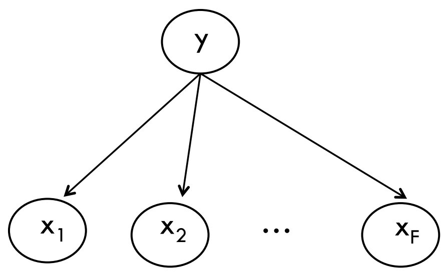
- Graphical models
  - Naïve Bayes classifiers
  - Bayesian networks
- Decision trees
  - Random forests (many decision trees)
- Neural Networks
  - Perceptrons
  - Artificial neural networks
  - Deep belief nets
- Max margin classifiers
  - Support vector machines
- Regression analysis
  - Logistic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

Each of these algorithms makes assumptions – these assumptions are known as the inductive bias of the classifier

Naïve Bayes Classifier

- Used for classification
  - x_i are symptoms and y = {Flu, Appendicitis,…}
  - x_i are word frequencies and y = {Politics, Sports, Finance,…}

- Inductive bias: features are conditionally independent given label

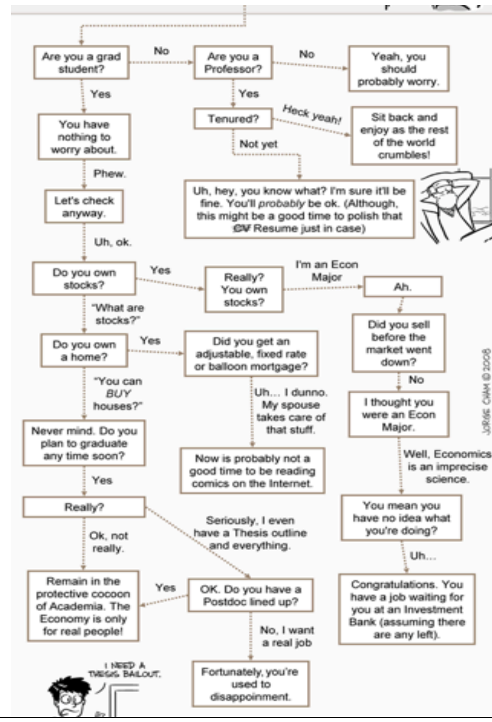


Naïve Bayes Classifier

- Training: learn p(y) and $p(x_f|y)$ from data set D
  - Think of D as a set of samples we observed
  - Use these samples to estimate distributions

- Testing: Once we estimate these probabilities from D, want to compute $p(y=k|x)$ for a new instance x
  - Assign x to whichever class has highest probability

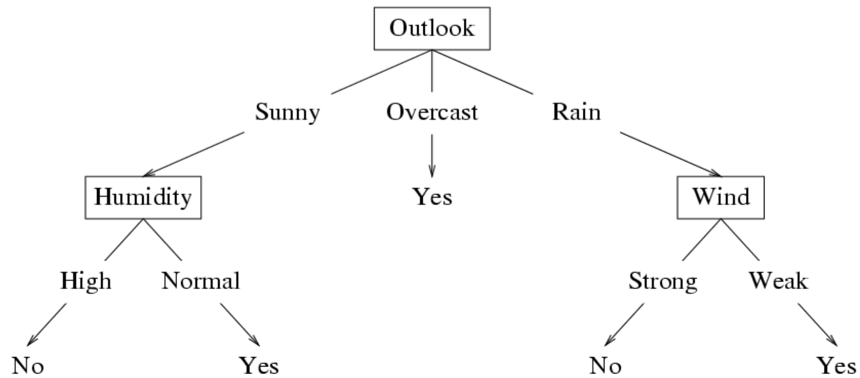
The Economic Meltdown:
Should you be concerned?
- PhD Comics



Decision Tree Classifier

	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 Tree Classifier

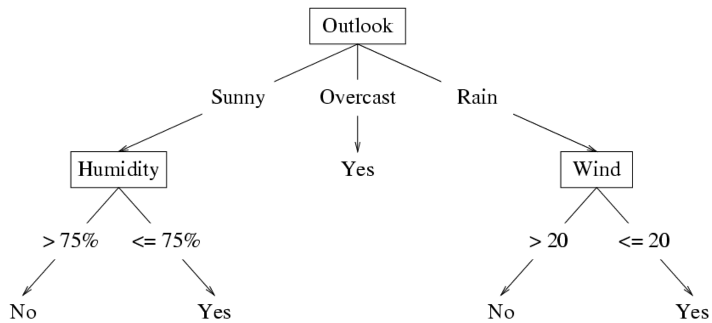


Decision Tree Classifier

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

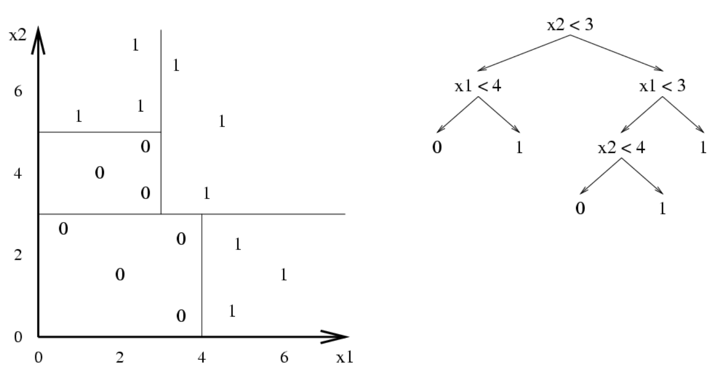
Decision Tree Classifier

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

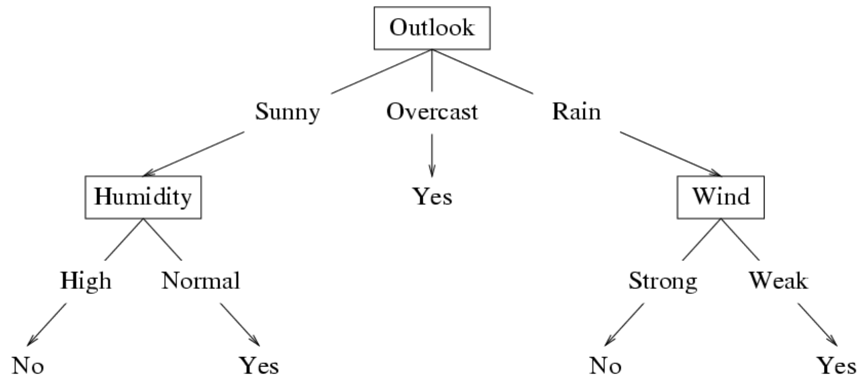


Decision Tree Classifier

- Learns axis-parallel decision boundaries, i.e. divides feature space into hyper-rectangles



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
```