SUPERVISED LEARNING

Progress Report

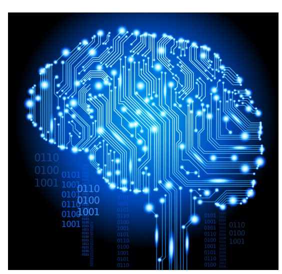
- We've finished Part I: Problem Solving
- We've finished Part II: Reasoning with uncertainty
- Part III: (Machine) Learning

Today

- 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

So what is learning?

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

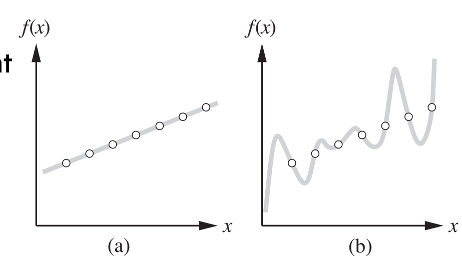
Types of Supervised Learning Tasks

- 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

Ockham's Razor

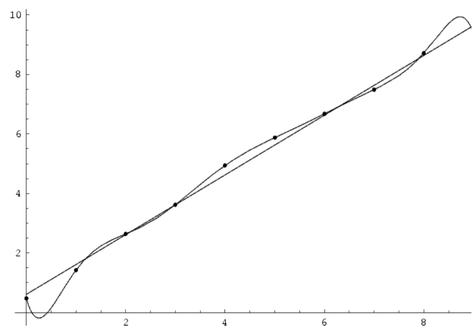
- Ockham's Razor
  - Prefer the simplest consistent hypothesis

- Example: Curve fitting
  - 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

- Inductive bias: features are conditionally independent given label

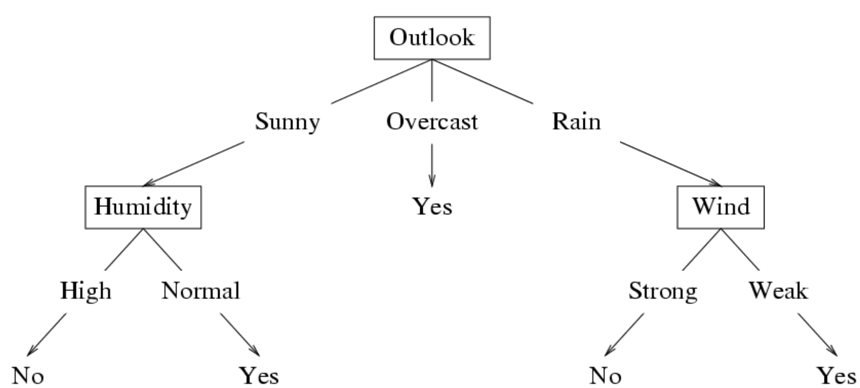
Naïve Bayes Classifier

- Training
- Testing

Decision Tree Classifier

	Day	Outlook	Temp.	Humidity	Wind	PlayTennis	
$\mathbf{x}_1 \rightarrow$	D1	Sunny	Hot	High	Weak	No	$\leftarrow y_1$
$\mathbf{x}_2 \rightarrow$	D2	Sunny	Hot	High	Strong	No	$\leftarrow y_2$
$\mathbf{x}_3 \rightarrow$	D3	Overcast	Hot	High	Weak	Yes	$\leftarrow y_3$
	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

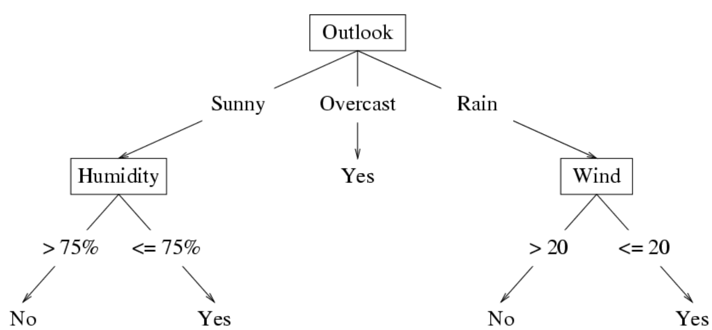


Decision Tree Classifier

- 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

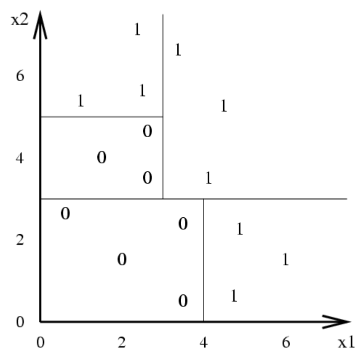
Decision Tree Classifier

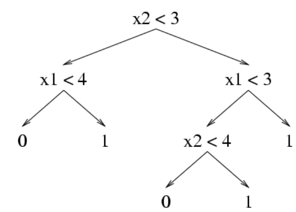
- If the features are continuous, internal nodes may test the value of a feature against a threshold



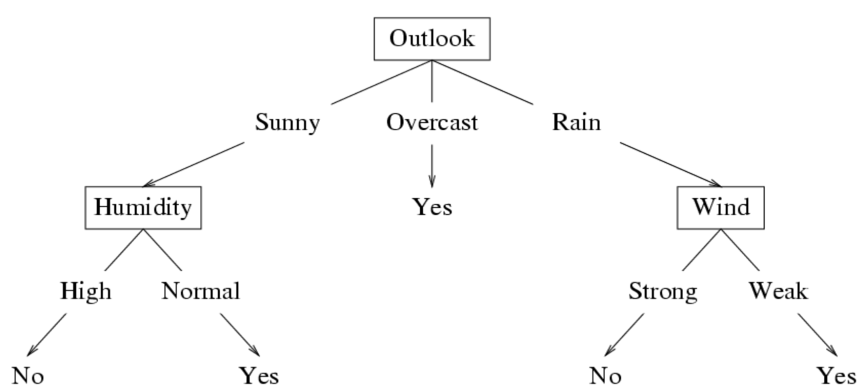
Decision Tree Classifier

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





Learning a Decision Tree



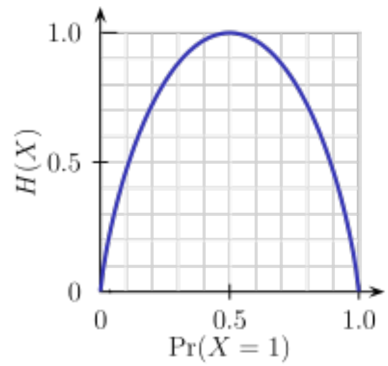
Pseudocode

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
- 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 applied to Decision Trees

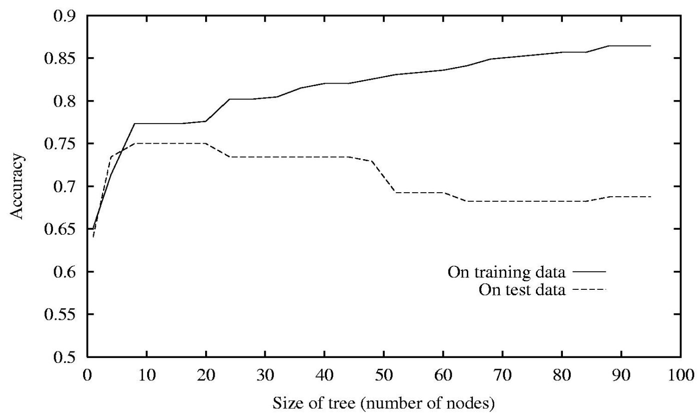
Entropy and Information Gain

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

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

Practice

Decision Trees: additional considerations



Decision Trees: additional considerations

- Overfitting
  - Can prune to improve performance on a validity set
- Continuous or integer-valued attributes
  - Use ranges
- Continuous label y
  - Combination of splitting and linear regression