Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 4, 2012

Today:
- What is machine learning?
- Decision tree learning
- Course logistics
- **Homework 1 handed out**

Readings:
- "The Discipline of ML"
- Mitchell, Chapter 3
- Bishop, Chapter 14.4

Machine Learning:

Study of algorithms that
- improve their performance P
- at some task T
- with experience E

well-defined learning task: <P,T,E>

Learning to Predict Emergency C-Sections

[Sims et al., 2000]

Data:

Patient103 time=1	→ *Patient103* time=2	... → *Patient103* time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: no	Elective C–Section: no
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
...	...	...

9714 patient records, each with 215 features

One of 18 learned rules:

```
If   No previous vaginal delivery, and
     Abnormal 2nd Trimester Ultrasound, and
     Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

 Over training data: 26/41 = .63,
 Over test data: 12/20 = .60
```

Learning to detect objects in images

(Prof. H. Schneiderman)

Example training images
for each orientation

Learning to classify text documents

→ Company home page

vs

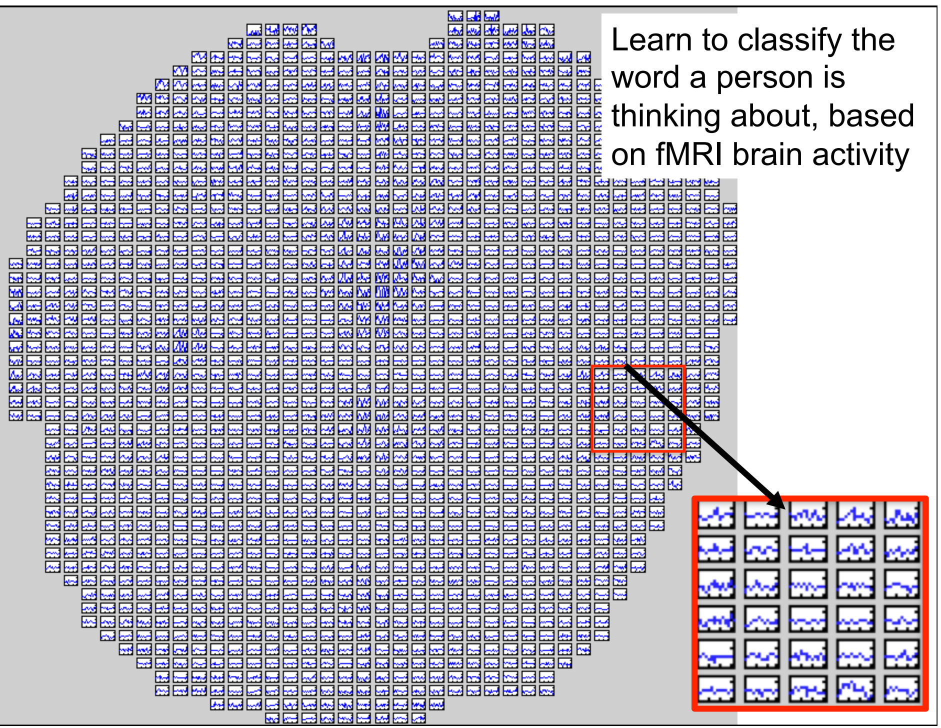
Personal home page

vs

University home page

vs

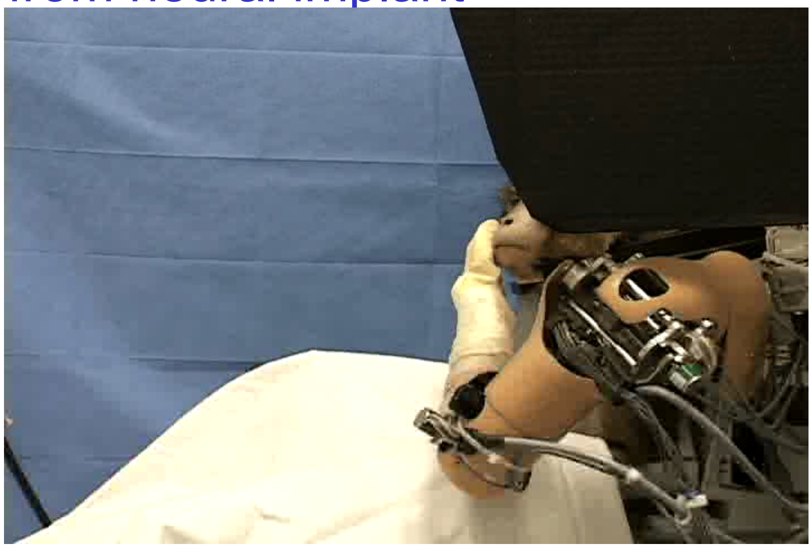
…



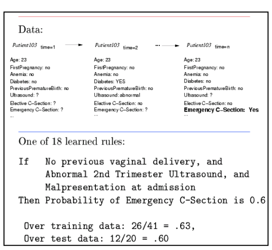
Learn to classify the word a person is thinking about, based on fMRI brain activity

Learning prosthetic control from neural implant

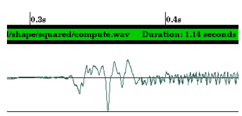
[R. Kass
L. Castellanos
A. Schwartz]



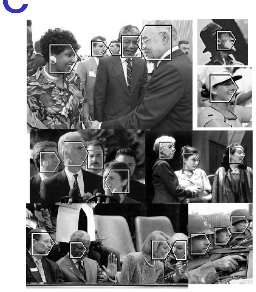
Machine Learning - Practice



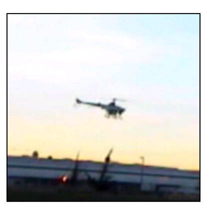
Mining Databases



Speech Recognition

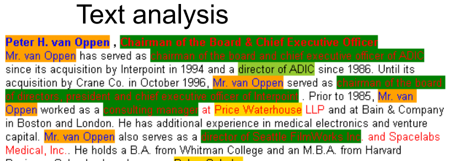


Object recognition



Control learning

Text analysis



- Supervised learning
- Bayesian networks
- Hidden Markov models
- Unsupervised clustering
- Reinforcement learning
- ....

Machine Learning - Theory

<u>PAC Learning Theory</u>
(supervised concept learning)

examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

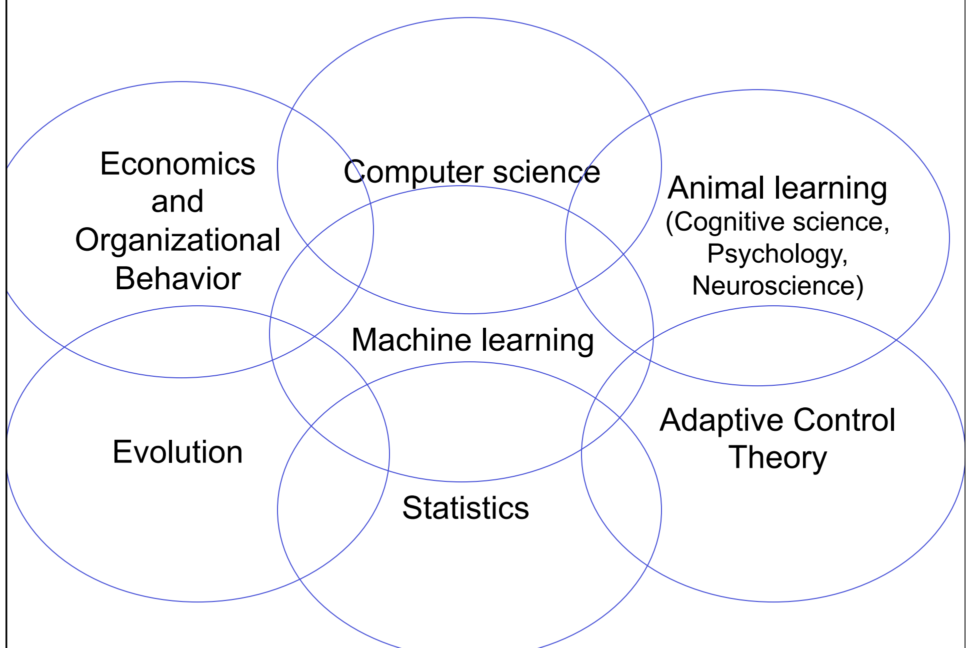
$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Other theories for

- Reinforcement skill learning
- Semi-supervised learning
- Active student querying
- …

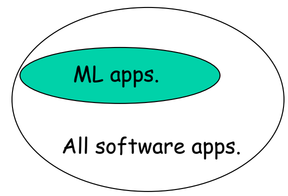
… also relating:

- # of mistakes during learning
- learner's query strategy
- convergence rate
- asymptotic performance
- bias, variance



Machine Learning in Computer Science

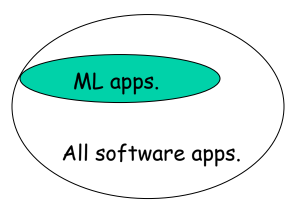
- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing (why?)

Machine Learning in Computer Science

- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing
  - Improved machine learning algorithms
  - Increased data capture, networking, new sensors
  - Software too complex to write by hand
  - Demand for self-customization to user, environment

Course logistics

Machine Learning 10-601

course page: www.cs.cmu.edu/~tom/10601_fall2012

Lecturers
- Ziv Bar-Joseph
- Tom Mitchell

TA's
- Brendan O'Conner
- Mehdi Samadi
- Selen Uguroglu
- Daegon Won

Course assistant
- Sharon Cavlovich (GHC 8215)

See webpage for
- Office hours
- Syllabus details
- Recitation sessions
- Grading policy
- Honesty policy
- Late homework policy
- ...

Highlights of Course Logistics

Recitation sessions:
- Optional, very helpful
- 5pm tues. and wed.
  - Duplicate sessions – pick one
- start TODAY
  - Matlab review Gates 6115

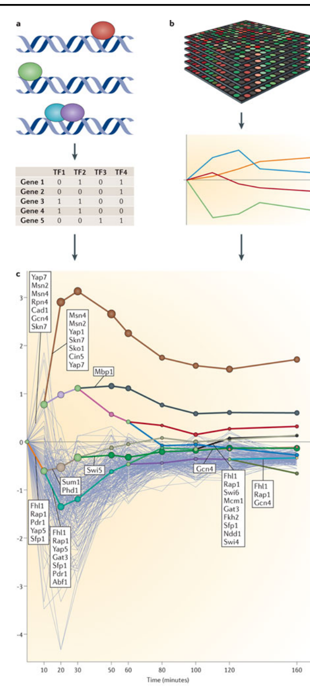
Grading:
- 40% homeworks (~5-6)
- 25% midterm
- 35% final exam

Late homework:
- full credit when due
- half credit next 48 hrs
- zero credit after that
- must turn in n-1 of the n homeworks, even if late

Being present at exams:
- You must be there – plan now.

Ziv Bar-Joseph

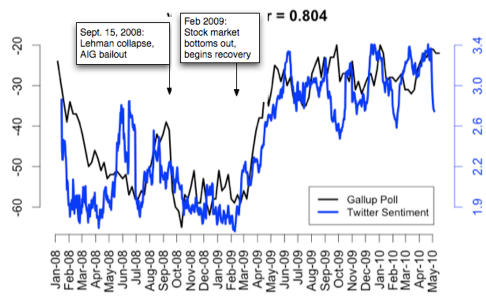
How can we integrate static and time series data to reconstruct dynamic models of biological systems?

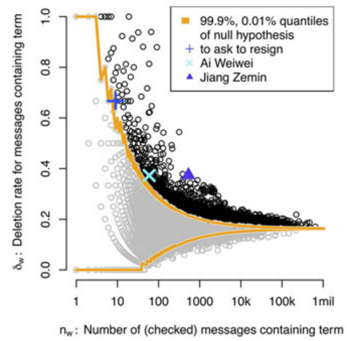


Brendan O'Connor

What can **statistical text analysis** tell us about **society**?
(tools for social science)

Twitter Sentiment and Polls
(they can correlate)





Chinese Microblog Censorship
(identify sensitive content)

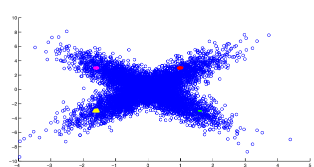
http://brenocon.com

Selen Uguroglu

Learning with rare classes

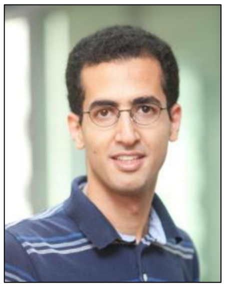
- Fraudulent credit card transactions
- Diagnosis of rare medical diseases
- Network intrusions

Active learning, feature selection when the dataset has highly skewed class distribution

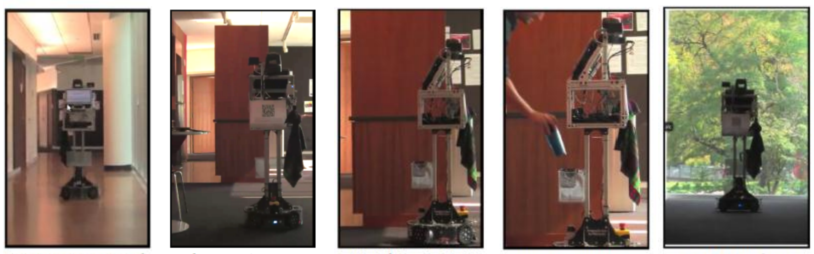


5th year graduate student in Language Technologies Institute (LTI), SCS
Homepage: www.cs.cmu.edu/~sugurogl

Mehdi Samadi

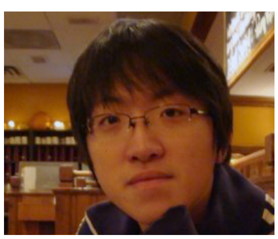


- Automate the combined retrieval and use of the underlying information on the Web.
- Extend the applicability of knowledge acquisition techniques for both automated agents and humans.

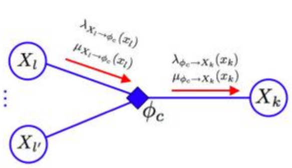


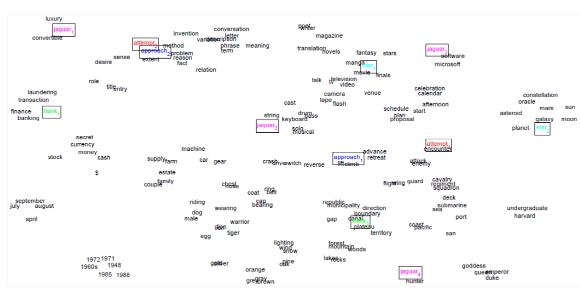
(a) Receive command (b) Go to location (c) Ask for "coffee" (d) Get "coffee" (e) Deliver object

Daegun Won



- Efficient inference method in graphical models
  - Incremental inference?
  - Degree of dependency?





- Past projects in
  - Active learning
  - Empirical phrasal synonym finding

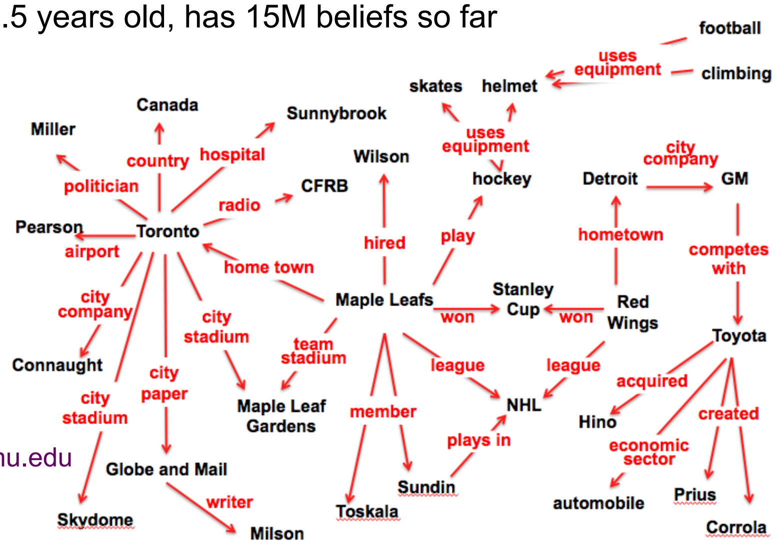
3rd year Ph.D. student at Language Technologies Institute
Homepage: www.cs.cmu.edu/~xichen

Tom Mitchell

How can we build never-ending learners?

NELL runs 24x7, learning to read the web

NELL now 2.5 years old, has 15M beliefs so far



http://rtw.ml.cmu.edu

Function Approximation and Decision tree learning

Function approximation

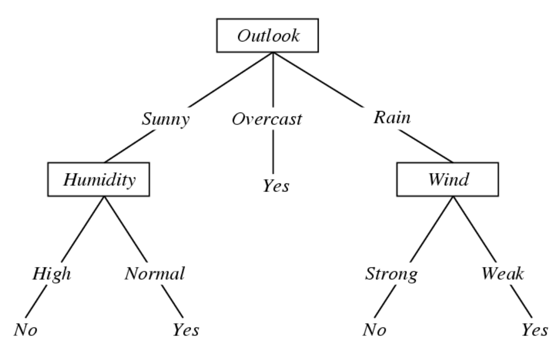
Problem Setting:
- Set of possible instances X
- Unknown target function $f : X \rightarrow Y$
- Set of function hypotheses $H=\{ h \mid h : X \rightarrow Y \}$

superscript: i^{th} training example

Input:
- Training examples $\{\langle x^{(i)},y^{(i)}\rangle\}$ of unknown target function f

Output:
- Hypothesis $h \in H$ that best approximates target function f

A Decision tree for
f: <Outlook, Humidity, Wind, Temp> $\rightarrow$ PlayTennis?



More generally, f: $\langle X_1, \ldots X_n \rangle \rightarrow Y$

Each internal node: discrete test on one attribute, X_i

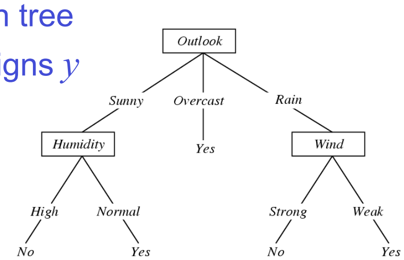
Each branch from a node: selects one value for X_i

Each leaf node: predict Y (or $P(Y|X \in \text{leaf})$)

Decision Tree Learning

Problem Setting:

- Set of possible instances X
  - each instance x in X is a feature vector
  - e.g., <*Humidity=low, Wind=weak, Outlook=rain, Temp=hot*>
- Unknown target function $f : X \rightarrow Y$
  - Y=1 if we play tennis on this day, else 0
- Set of function hypotheses H={ h | $h : X \rightarrow Y$ }
  - each hypothesis h is a decision tree
  - trees sorts x to leaf, which assigns y



Decision Tree Learning

Problem Setting:

- Set of possible instances X
  - each instance x in X is a feature vector
    $x = < x_1, x_2 \ldots x_n >$
- Unknown target function $f : X \rightarrow Y$
  - Y is discrete-valued
- Set of function hypotheses H={ h | $h : X \rightarrow Y$ }
  - each hypothesis h is a decision tree

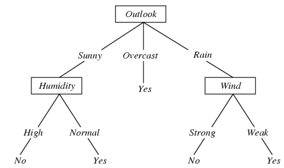
Input:

- Training examples {$<x^{(i)},y^{(i)}>$} of unknown target function f

Output:

- Hypothesis $h \in H$ that best approximates target function f

Decision Trees



Suppose $X = <X_1, \ldots X_n>$
where X_i are boolean variables

How would you represent $Y = X_2 X_5$? $Y = X_2 \vee X_5$

How would you represent $X_2 X_5 \vee X_3 X_4 (\neg X_1)$

A Tree to Predict C-Section Risk

Learned from medical records of 1000 women

Negative examples are C-sections

```
[833+,167-] .83+ .17-
Fetal_Presentation = 1: [822+,116-] .88+ .12-
| Previous_Csection = 0: [767+,81-] .90+ .10-
| | Primiparous = 0: [399+,13-] .97+ .03-
| | Primiparous = 1: [368+,68-] .84+ .16-
| | | Fetal_Distress = 0: [334+,47-] .88+ .12-
| | | | Birth_Weight < 3349: [201+,10.6-] .95+ .
| | | | Birth_Weight >= 3349: [133+,36.4-] .78+
| | | Fetal_Distress = 1: [34+,21-] .62+ .38-
| Previous_Csection = 1: [55+,35-] .61+ .39-
Fetal_Presentation = 2: [3+,29-] .11+ .89-
Fetal_Presentation = 3: [8+,22-] .27+ .73-
```

Top-Down Induction of Decision Trees

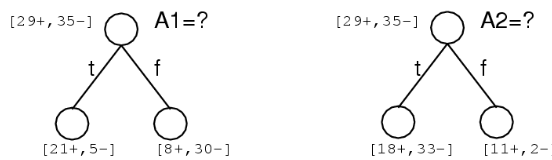
[ID3, C4.5, Quinlan]

node = Root

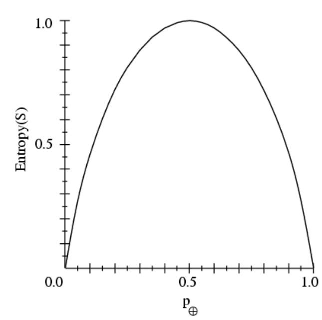
Main loop:

1. $A \leftarrow$ the "best" decision attribute for next *node*
2. Assign A as decision attribute for *node*
3. For each value of A, create new descendant of *node*
4. Sort training examples to leaf nodes
5. If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes

Which attribute is best?



Sample Entropy



- S is a sample of training examples
- $p_{\oplus}$ is the proportion of positive examples in S
- $p_{\ominus}$ is the proportion of negative examples in S
- Entropy measures the impurity of S

$$H(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Entropy

Entropy $H(X)$ of a random variable X

# of possible values for X

$$H(X) = -\sum_{i=1}^{n} P(X=i) \log_2 P(X=i)$$

$H(X)$ is the expected number of bits needed to encode a randomly drawn value of X (under most efficient code)

Why? Information theory:

- Most efficient possible code assigns $-\log_2 P(X=i)$ bits to encode the message $X=i$
- So, expected number of bits to code one random X is:

$$\sum_{i=1}^{n} P(X=i)(-\log_2 P(X=i))$$

Entropy

Entropy $H(X)$ of a random variable X

$$H(X) = -\sum_{i=1}^{n} P(X=i) \log_2 P(X=i)$$

Specific conditional entropy $H(X|Y=v)$ of X given $Y=v$:

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v) \log_2 P(X=i|Y=v)$$

Conditional entropy $H(X|Y)$ of X given Y :

$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v) H(X|Y=v)$$

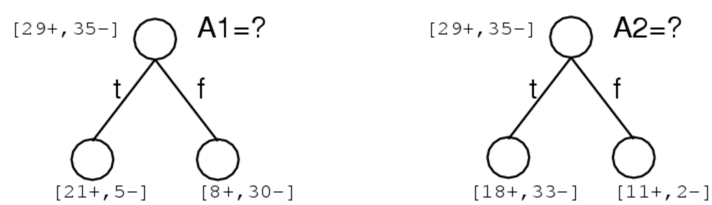
Mutual information (aka Information Gain) of X and Y :

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Information Gain is the mutual information between input attribute A and target variable Y

Information Gain is the expected reduction in entropy of target variable Y for data sample S, due to sorting on variable A

$$Gain(S, A) = I_S(A, Y) = H_S(Y) - H_S(Y|A)$$

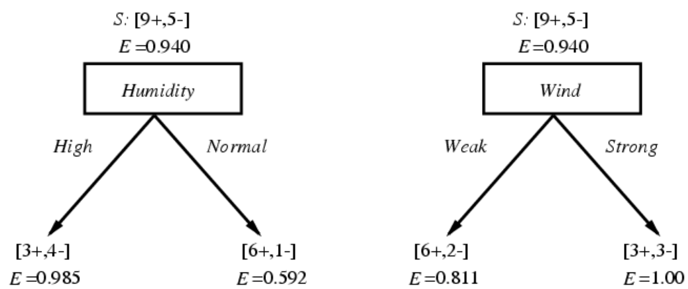


Training Examples

Day	Outlook	Temperature	Humidity	Wind	PlayTen
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

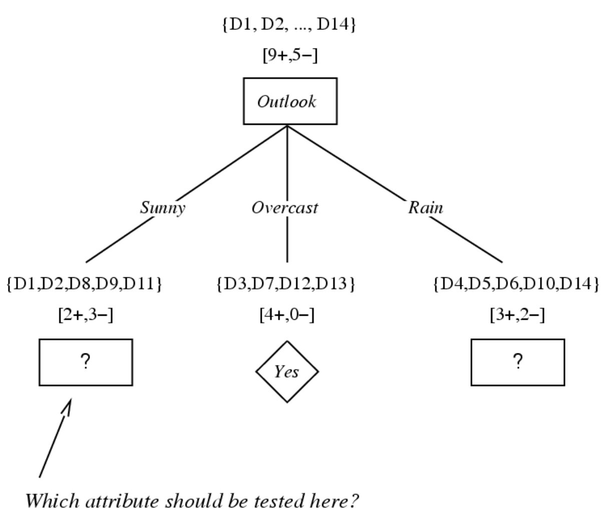
Selecting the Next Attribute

Which attribute is the best classifier?



$Gain(S, Humidity)$
$= .940 - (7/14).985 - (7/14).592$
$= .151$

$Gain(S, Wind)$
$= .940 - (8/14).811 - (6/14)1.0$
$= .048$



$S_{sunny} = \{D1,D2,D8,D9,D11\}$

$Gain(S_{sunny}, Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 = .970$

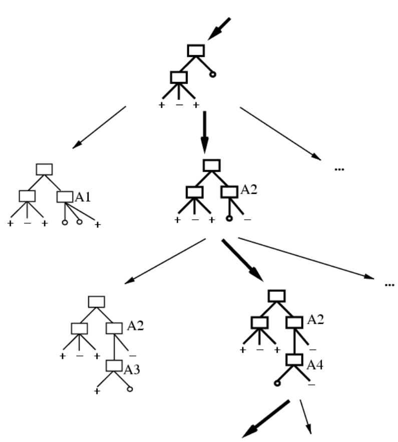
$Gain(S_{sunny}, Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$

$Gain(S_{sunny}, Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$

Decision Tree Learning Applet

- http://www.cs.ualberta.ca/%7Eaixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html

Which Tree Should We Output?



- ID3 performs heuristic search through space of decision trees
- It stops at smallest acceptable tree. Why?

Occam's razor: prefer the simplest hypothesis that fits the data

Why Prefer Short Hypotheses? (Occam's Razor)

Arguments in favor:

Arguments opposed:

Why Prefer Short Hypotheses? (Occam's Razor)

Argument in favor:

- Fewer short hypotheses than long ones

→ a short hypothesis that fits the data is less likely to be a statistical coincidence

→ highly probable that a sufficiently complex hypothesis will fit the data

Argument opposed:

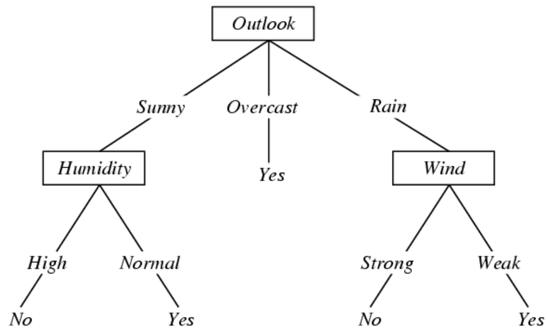
- Also fewer hypotheses with prime number of nodes and attributes beginning with "Z"
- What's so special about "short" hypotheses?

Overfitting in Decision Trees

Consider adding noisy training example #15:

$Sunny,\ Hot,\ Normal,\ Strong,\ PlayTennis = No$

What effect on earlier tree?



Overfitting

Consider a hypothesis h and its

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

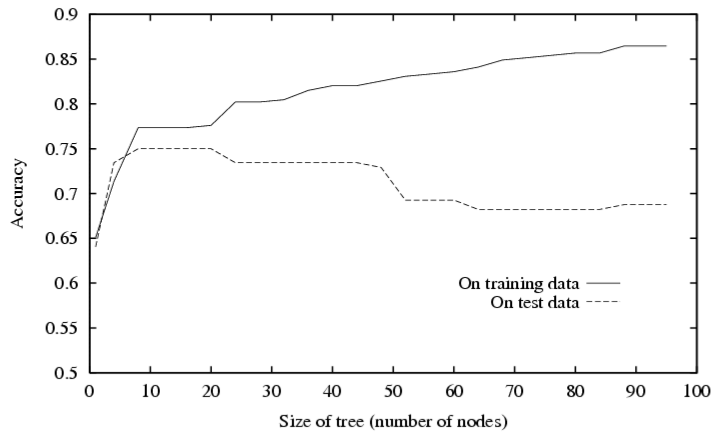
We say h overfits the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

Overfitting in Decision Tree Learning



Avoiding Overfitting

How can we avoid overfitting?

- stop growing when data split not statistically significant
- grow full tree, then post-prune

Reduced-Error Pruning

Split data into *training* and *validation* set

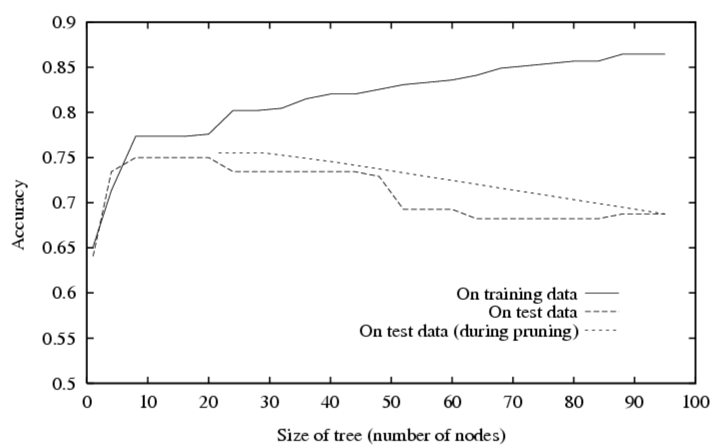
Create tree that classifies *training* set correctly

Do until further pruning is harmful:

1. Evaluate impact on *validation* set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves *validation* set accuracy

- produces smallest version of most accurate subtree
- What if data is limited?

Effect of Reduced-Error Pruning



Continuous Valued Attributes

Create a discrete attribute to test continuous

- $Temperature = 82.5$
- $(Temperature > 72.3) = t, f$

Temperature:	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No

Attributes with Many Values

Problem:

- If attribute has many values, $Gain$ will select it
- Imagine using $Date = Jun\_3\_1996$ as attribute

One approach: use $GainRatio$ instead

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

$$SplitInformation(S, A) \equiv - \sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

where S_i is subset of S for which A has value v_i

You should know:

- Well posed function approximation problems:
  - Instance space, X
  - Sample of labeled training data { <$x^{(i)}$, $y^{(i)}$>}
  - Hypothesis space, H = { f: X→Y }
- Learning is a search/optimization problem over H
  - Various objective functions
    - minimize training error (0-1 loss)
    - among hypotheses that minimize training error, select smallest (?)
- Decision tree learning
  - Greedy top-down learning of decision trees (ID3, C4.5, ...)
  - Overfitting and tree/rule post-pruning
  - Extensions…

Questions to think about (1)

- ID3 and C4.5 are heuristic algorithms that search through the space of decision trees. Why not just do an exhaustive search?

Questions to think about (2)

- Consider target function f: <x1,x2> → y, where x1 and x2 are real-valued, y is boolean. What is the set of decision surfaces describable with decision trees that use each attribute at most once?

Questions to think about (3)

- Why use Information Gain to select attributes in decision trees? What other criteria seem reasonable, and what are the tradeoffs in making this choice?

Questions to think about (4)

- What is the relationship between learning decision trees, and learning IF-THEN rules

One of 18 learned rules:

```
If   No previous vaginal delivery, and
     Abnormal 2nd Trimester Ultrasound, and
     Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

 Over training data: 26/41 = .63,
 Over test data: 12/20 = .60
```