Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 13, 2011

Today:
- What is machine learning?
- Decision tree learning
- Course logistics

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:

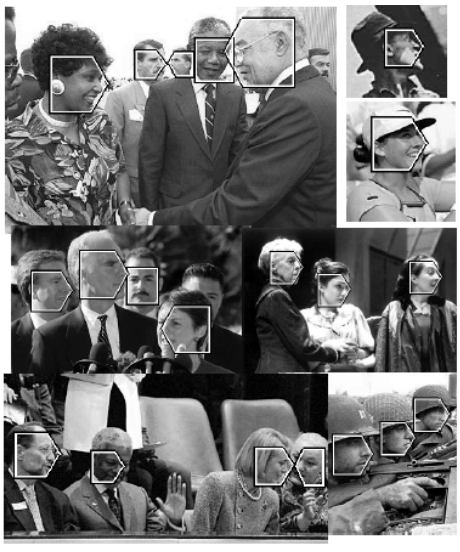
```
If   No previous vaginal delivery, and
     Abnormal 2nd Trimester Ultrasound, and
     Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

 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

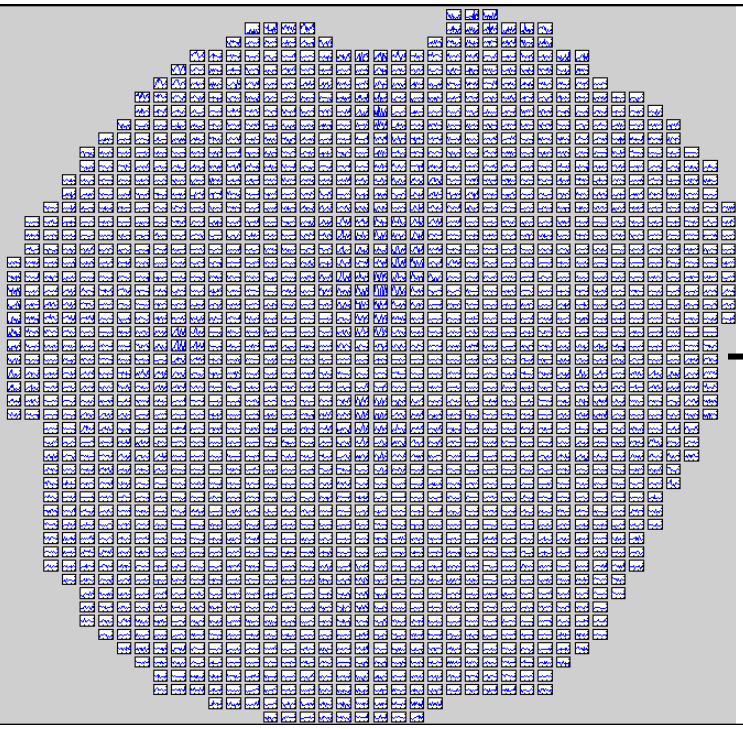
Personal home page

vs

University home page

vs

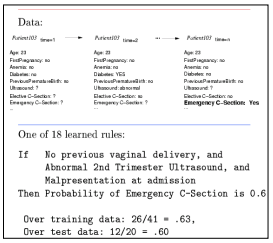
…



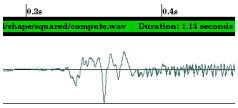
→ Reading a noun (vs verb)

[Rustandi et al., 2005]

Machine Learning - Practice

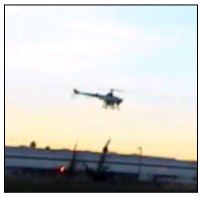


Mining Databases



Speech Recognition

Object recognition



Control learning

Text analysis

Peter H. van Oppen , Chairman of the Board & Chief Executive Officer

Mr. van Oppen has served as Chairman of the board and chief executive officer of ADIC since its acquisition by Interpoint in 1994 and a director of ADIC since 1986. Until its acquisition by Crane Co. in October 1996, Mr. van Oppen served as Chairman of the board of directors, president and chief executive officer of Interpoint . Prior to 1985, Mr. van Oppen worked as a consulting manager at Price Waterhouse LLP and at Bain & Company in Boston and London. He has additional experience in medical electronics and venture capital. Mr. van Oppen also serves as a director of Seattle FilmWorks Inc. and Spacelabs Medical, Inc.. He holds a B.A. from Whitman College and an M.B.A. from Harvard Business School, where he was a Baker Scholar.

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

Machine Learning - Theory

PAC Learning Theory
(supervised concept learning)

examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

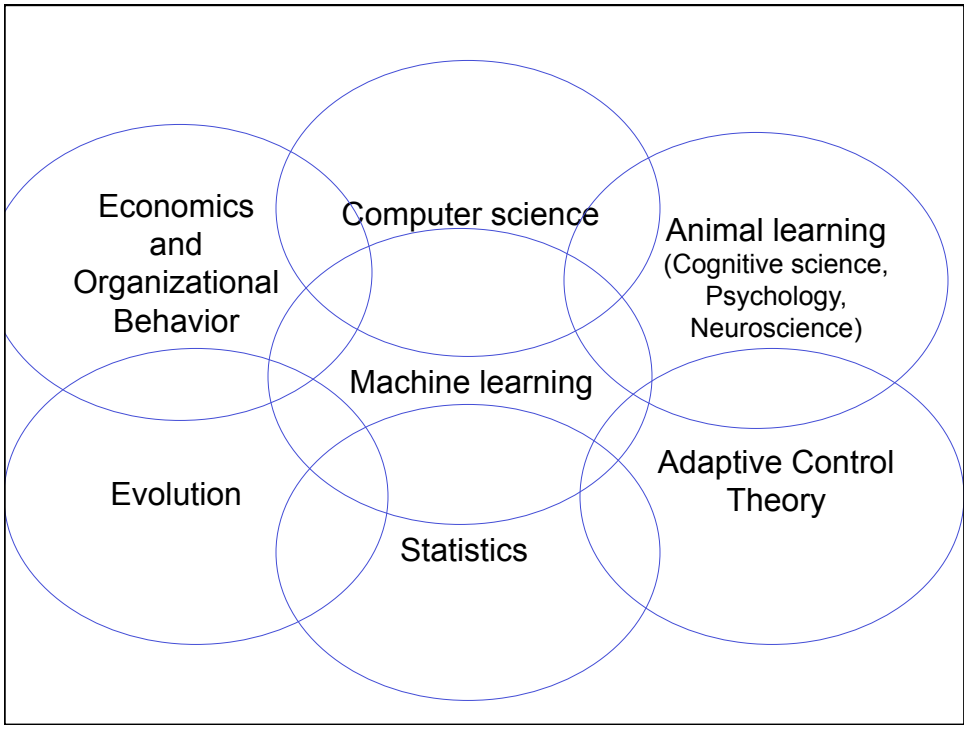
$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Other theories for

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

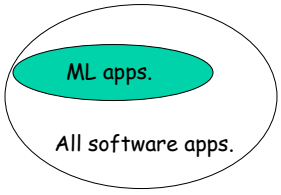
… also relating:

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



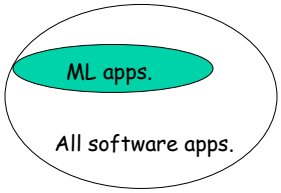
Machine Learning in Computer Science

- 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

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 $H=\{ h \mid h : X \rightarrow Y \}$

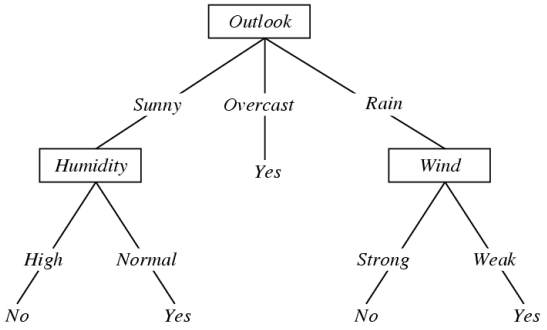
superscript: i^{th} training example

Input:
- Training examples $\{<x^{(i)}, y^{(i)}>\}$ 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> → PlayTennis?



Each internal node: test one discrete-valued attribute X_i

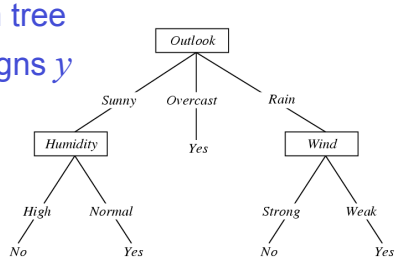
Each branch from a node: selects one value for X_i

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 is discrete valued
- Set of function hypotheses $H=\{ h \mid h : X \rightarrow Y \}$
  - 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 $H=\{ h \mid h : X \rightarrow Y \}$
  - each hypothesis h is a decision tree

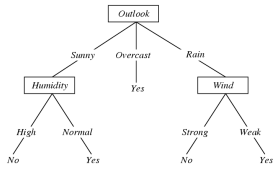
Input:

- 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]

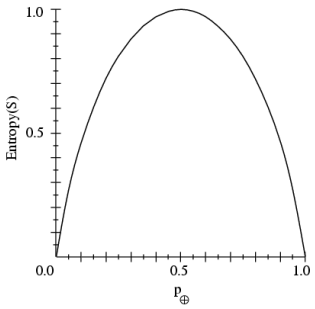
node = Root

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 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 :

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

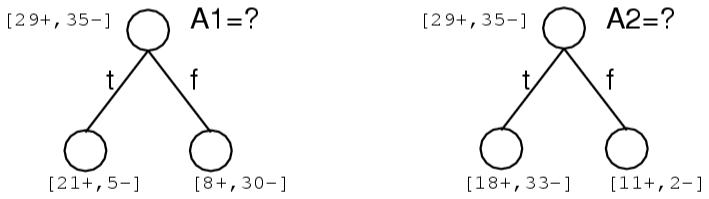
Mututal information (aka Information Gain) of X and Y :

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

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

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

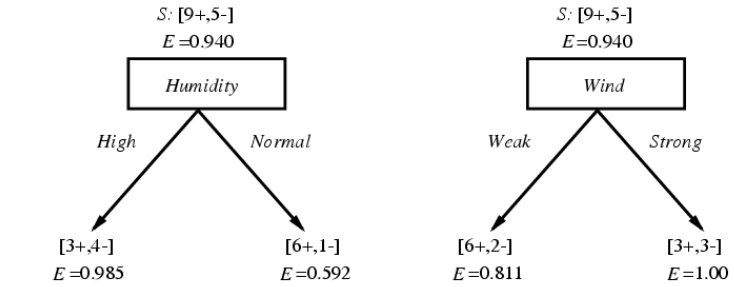


Training Examples

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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

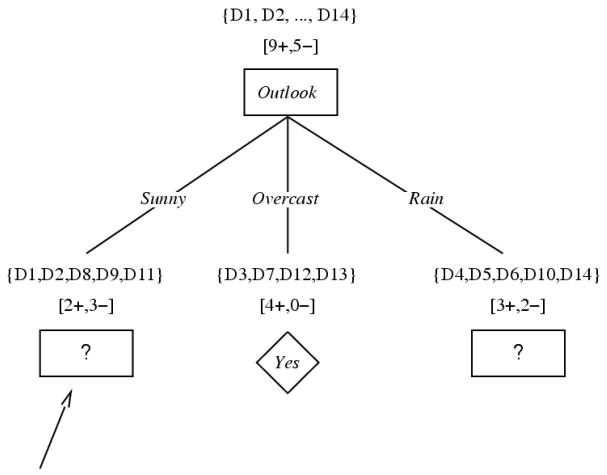
Selecting the Next Attribute

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



Which attribute should be tested here?

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

Gain (S_{sunny}, *Humidity*) = .970 − (3/5) 0.0 − (2/5) 0.0 = .970

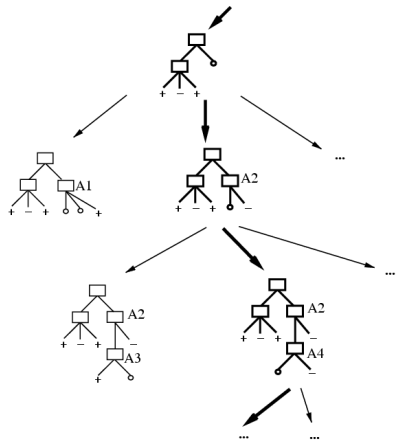
Gain (S_{sunny}, *Temperature*) = .970 − (2/5) 0.0 − (2/5) 1.0 − (1/5) 0.0 = .570

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:

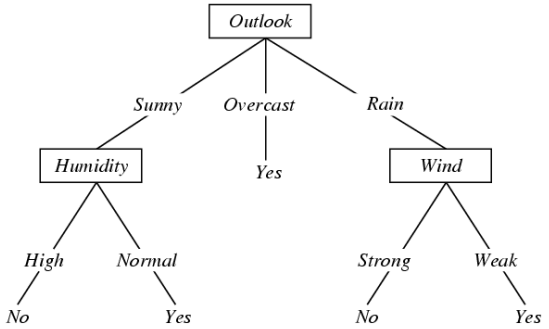
- Also fewer hypotheses with prime number of nodes and attributes beginning with "Z"
- What's so special about "short" hypotheses?

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: $error_{true}(h)$

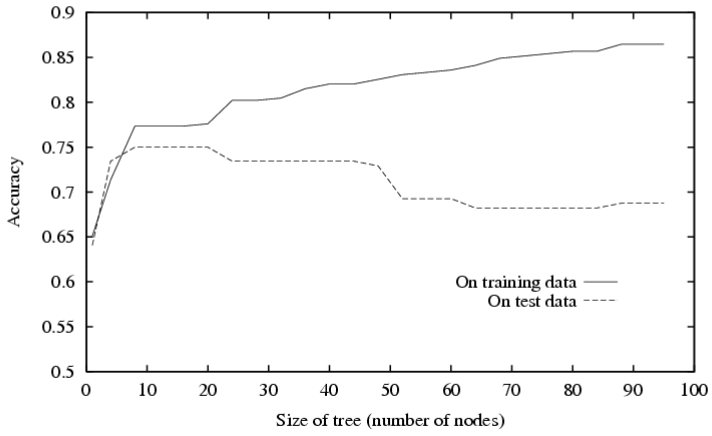
We say h overfits the training data if

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

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

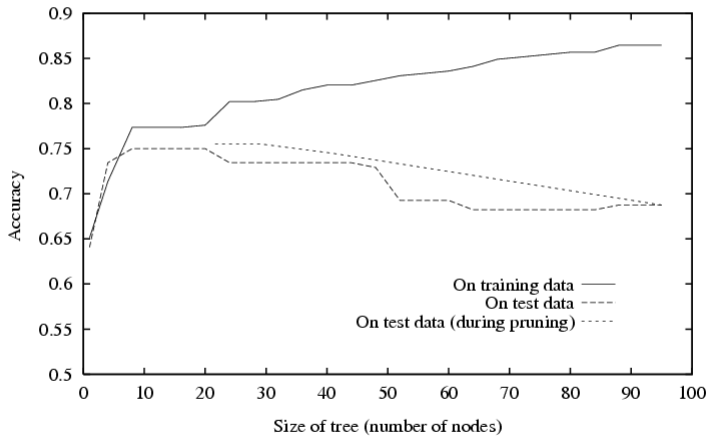
Create tree that classifies *training* set correctly

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

What 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?

Course logistics

Machine Learning 10-601

course page: www.cs.cmu.edu/~aarti/Class/10601

Lecturers
- Aarti Singh
- Tom Mitchell

TA's
- Will Bishop
- Shing-Hon Lau
- Mladen Kolar

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
- 5 or 6pm, wednesdays, (depending on room)
- start THIS WEEK. (watch email: possibly 6pm wed.)

Grading:

- 30% homeworks (~5)
- 25% midterm (October 27)
- 25% final exam (date tba)
- 20% course project

Late homework:

- full credit when due
- half credit next 48 hrs
- zero credit after that
- we will delete your lowest HW score
- must turn in n-1 of the n homeworks, even if late

Being present at exams:

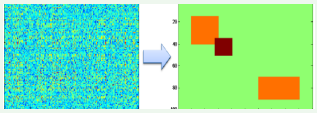
- You must be – plan now

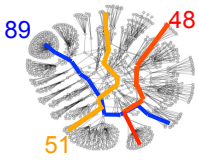
Aarti Singh

www.cs.cmu.edu/~aarti

Learning in **high-dimensional** systems using **corrupt**, **partial**, **compressed** and **active** measurements

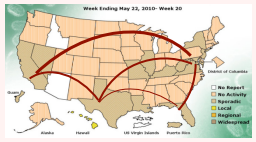
Robust and efficient Clustering in high-dimensions





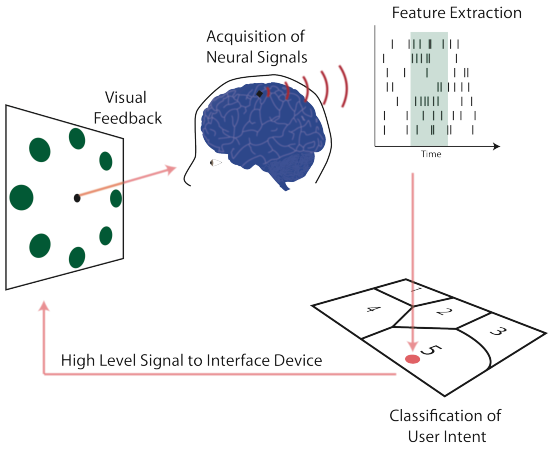
Learning large graph structures using few, selective measurements

Identifying weak patterns in networks - detecting nascent epidemics using non-local connectivity (**open position: undergrad**)



Will Bishop

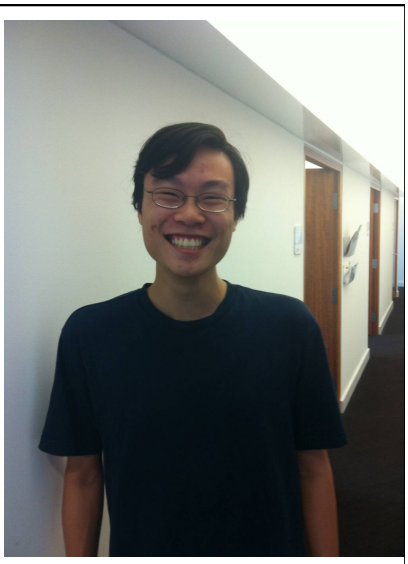
Online, semi-supervised learning for brain-computer interface: <u>Can we develop classifiers of brain signals that "autonomously" improve their performance over time?</u>

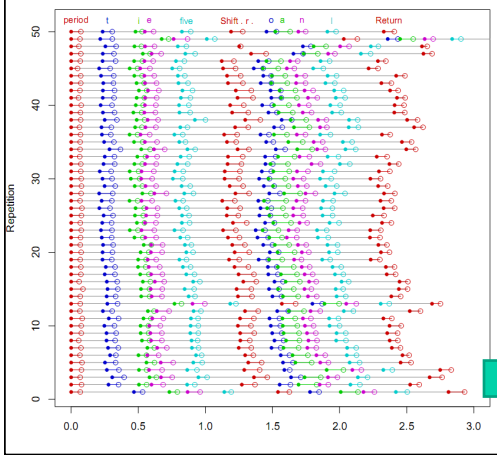


Shing-hon Lau

How do we conduct rigorous experiments in computer security?

Can we use the way a person types as a digital fingerprint?



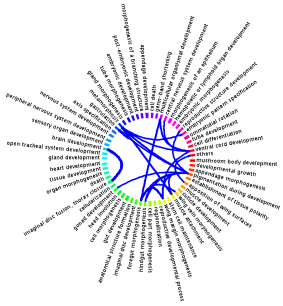


"Welcome back, Bob Smith."

Mladen Kolar

How to learn when the number of parameters is much larger than the number of sample points?
•How do we exploit unknown underlying structure?

Can we learn the structure?
•Sparsity
•Manifolds
•Low rank matrices
•…



5th year Ph.D. student at Machine Learning Dept. @ CMU. Advisor: Eric P. Xing.
Homepage: http://sites.coffeejunkies.org/mkolar/home

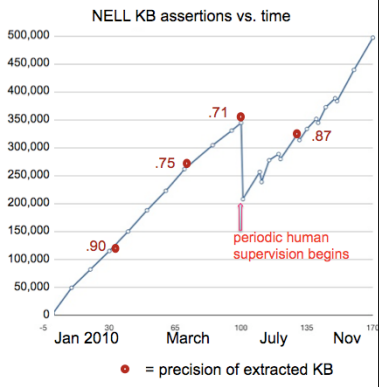
Tom Mitchell

How can we build never-ending learners?

Case study: never-ending language learner (NELL) runs 24x7 to learn to read the web

Recently-Learned Facts twitter

instance
tony_hall is a U.S._politician
christmas_lights is a perception_event
lung_carcinomas is a disease
wild_cat_beach is a beach
mike_epps is a commedian
ernest_hemingway contributed_to the creative work a_farewell_to_arms
andrew_stanton directed the movie finding_nemo
jeffords is a politician who holds the office of senators
time_warner controls cnn
the location of davis is california



see http://rtw.ml.cmu.edu