



# Machine Learning 10-601

Tom M. Mitchell  
Machine Learning Department  
Carnegie Mellon University

January 14, 2015

## Today:

- The Big Picture
- Overfitting
- Review: probability

## Readings:

- Decision trees, overfitting
- Mitchell, Chapter 3

## Probability review

- Bishop Ch. 1 thru 1.2.3
- Bishop, Ch. 2 thru 2.2
- Andrew Moore's online tutorial

# Function Approximation:

## **Problem Setting:**

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

## **Input:**

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

## **Output:**

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

# Function Approximation: Decision Tree Learning

## Problem Setting:

- Set of possible instances  $X$ 
  - each instance  $x$  in  $X$  is a feature vector  
 $x = \langle x_1, x_2 \dots x_n \rangle$
- 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  $\{ \langle x^{(i)}, y^{(i)} \rangle \}$  of unknown target function  $f$

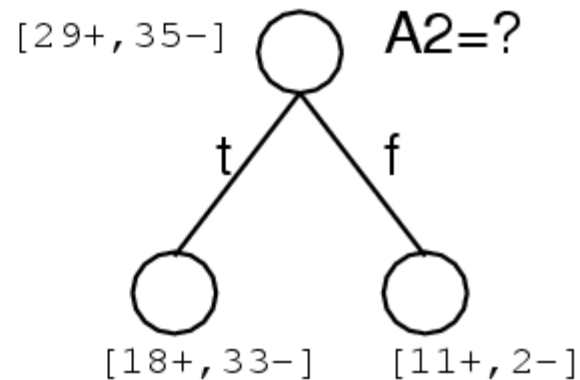
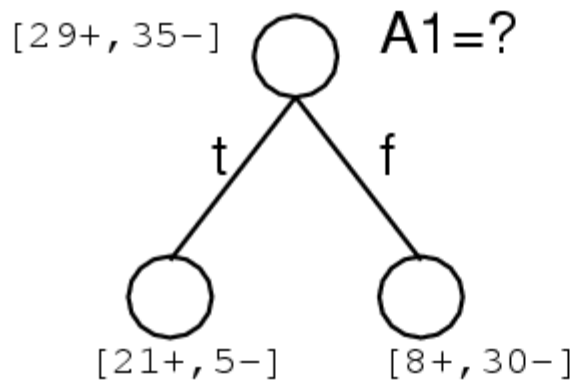
## Output:

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

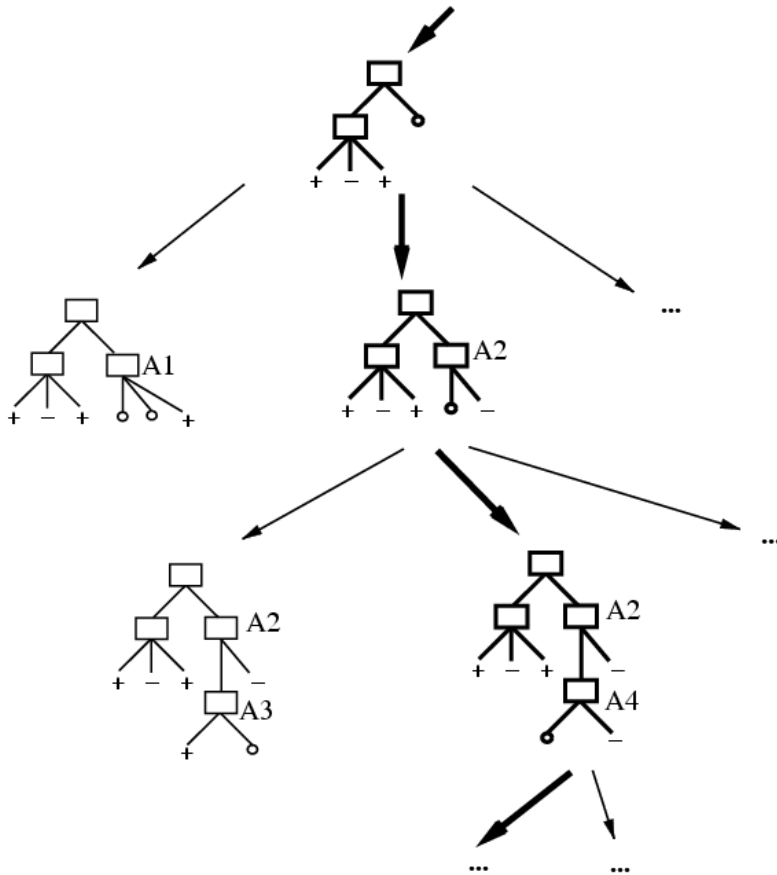
Information Gain (also called mutual information)  
between input attribute A and target variable Y

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

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



# Function approximation as Search for the best hypothesis

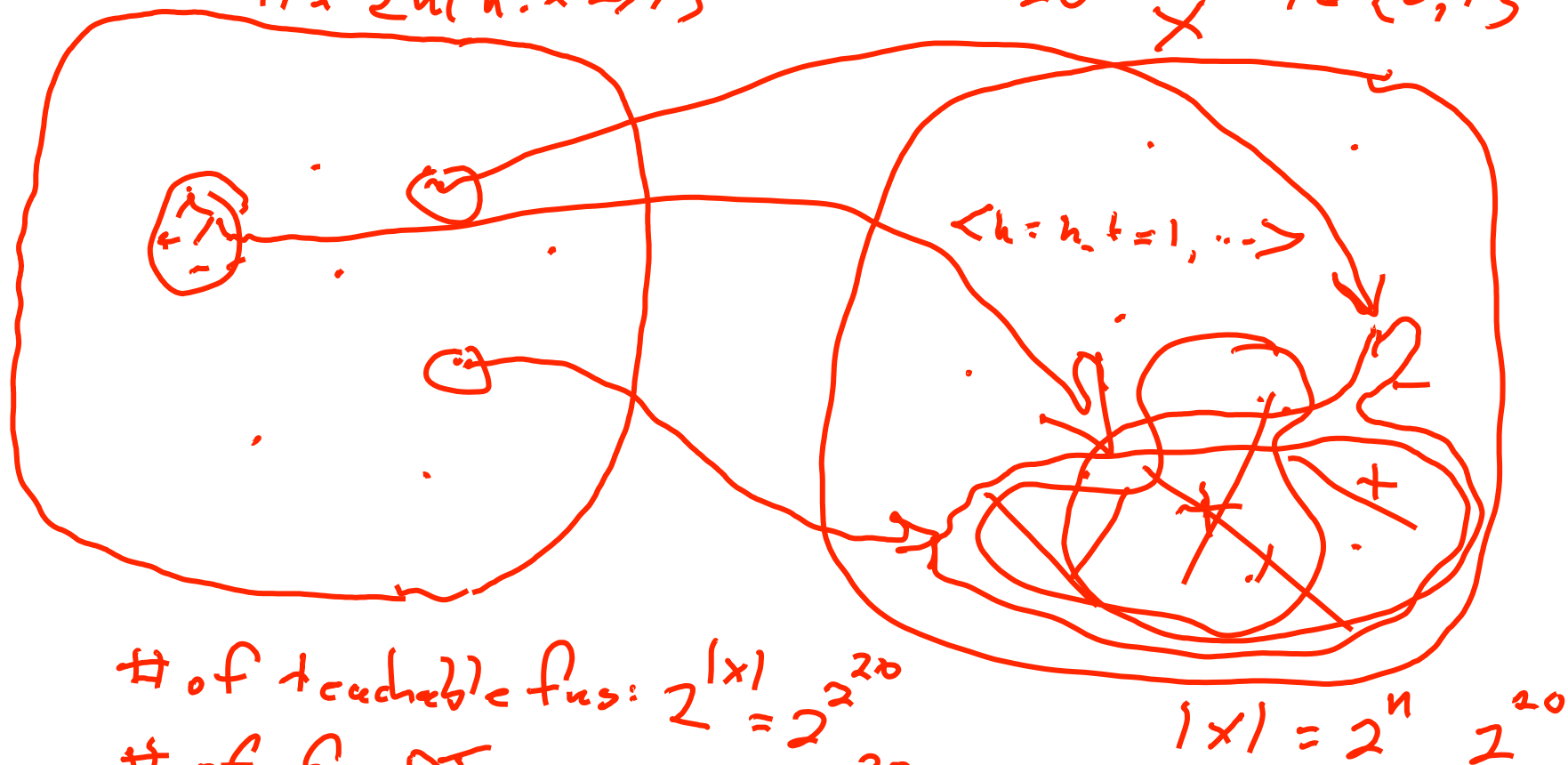


- ID3 performs heuristic search through space of decision trees

# Function Approximation: The Big Picture

$$f: x \rightarrow y \quad x = \langle x_1, \dots, x_n \rangle \quad x_i \in \{0, 1\}$$

$$H = \{h \mid h: x \rightarrow y\} \quad \uparrow_{20} \quad y \in \{0, 1\}$$



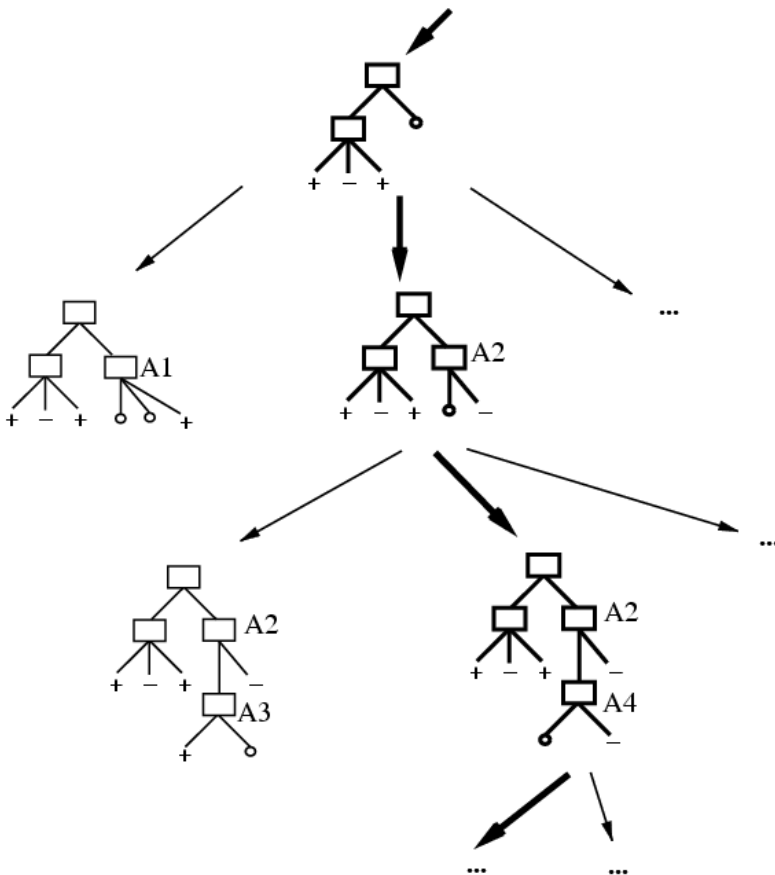
$$\# \text{ of teachable fns: } 2^{|X|} = 2^{2^{20}}$$

$$\# \text{ of fns DT's can repr} = 2^{2^{20}}$$

No free lunch!

# 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

Argument opposed:

- Also fewer hypotheses containing a prime number of nodes and attributes beginning with “Z”
- What's so special about “short” hypotheses, instead of “prime number of nodes and edges”?

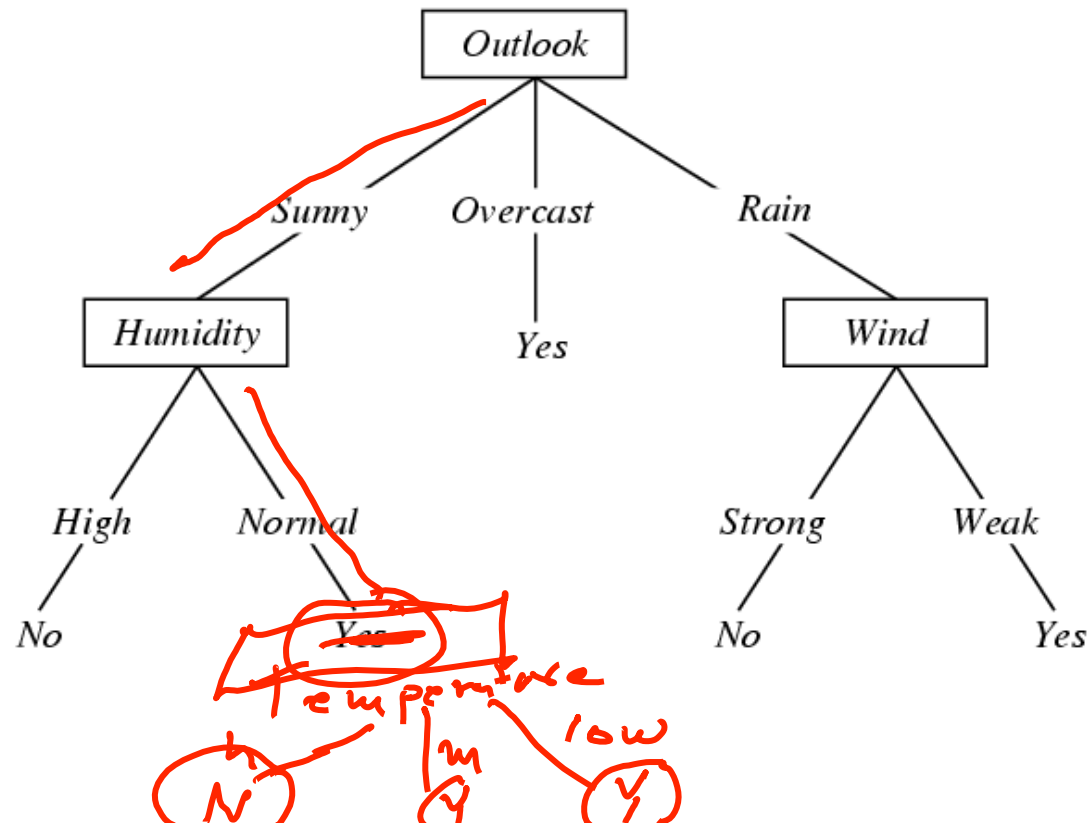
# Overfitting in Decision Trees

---

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:  $error_{true}(h)$



# 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

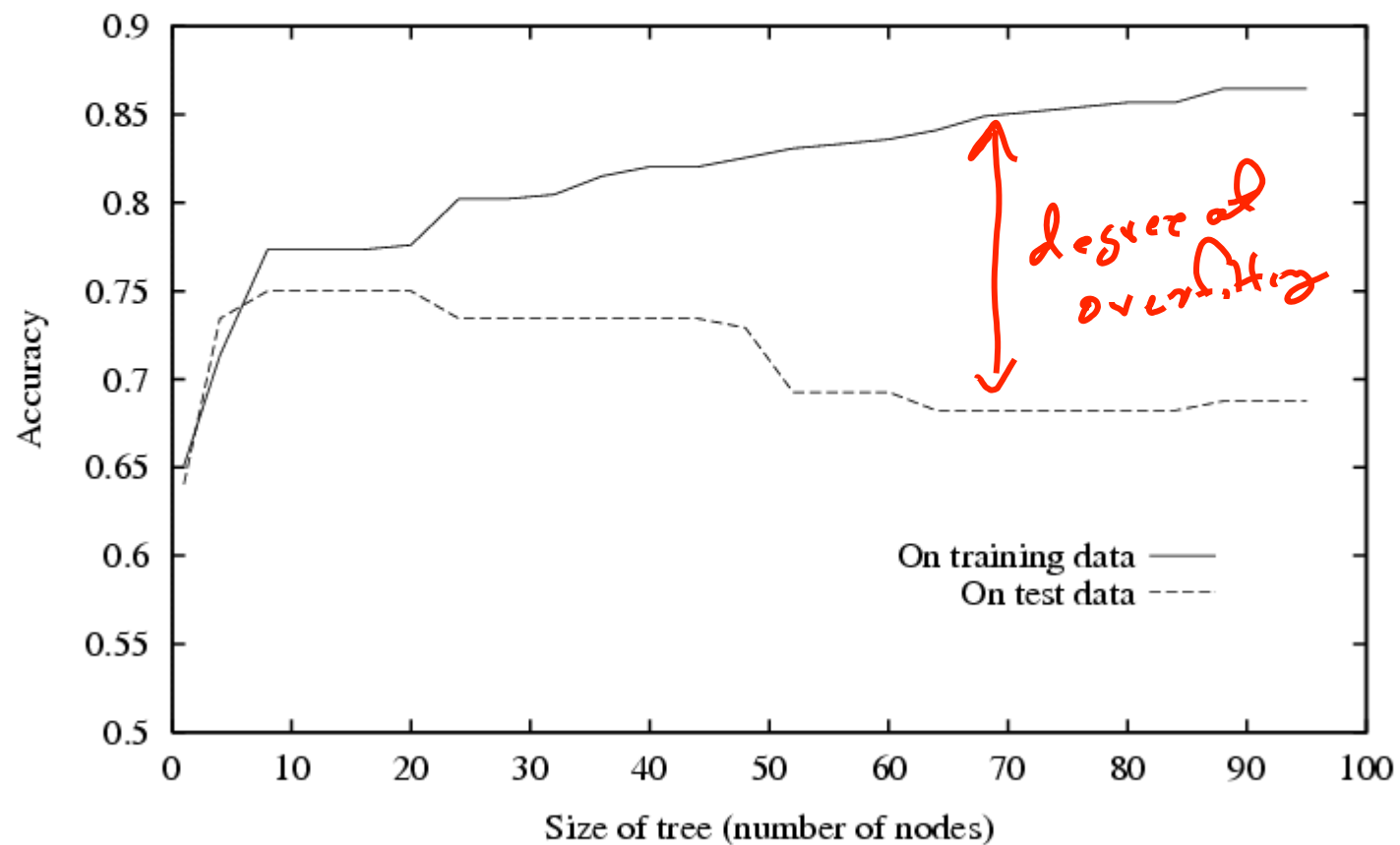
$$error_{true}(h) > error_{train}(h)$$

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

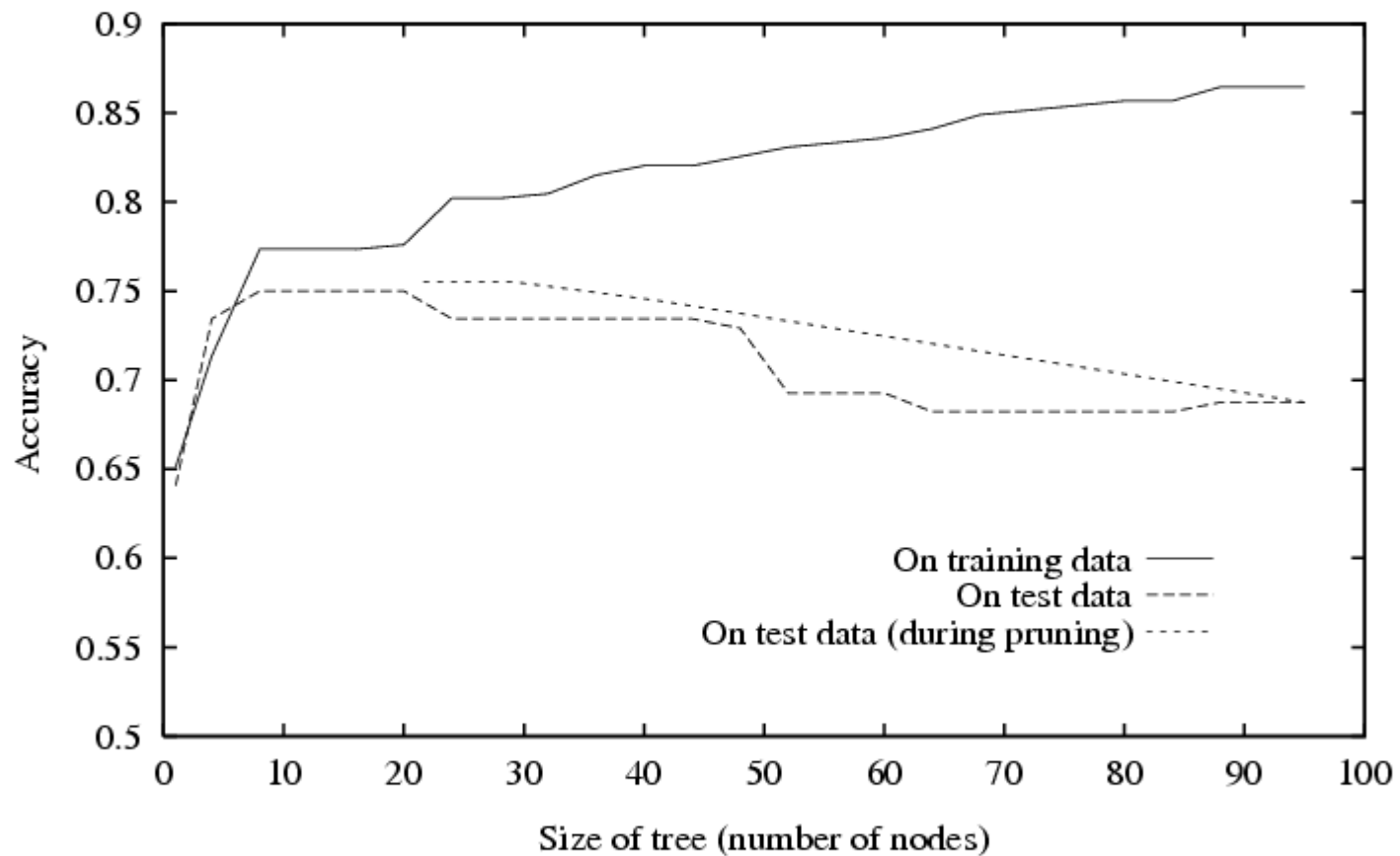
Create tree that classifies *training* set correctly

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

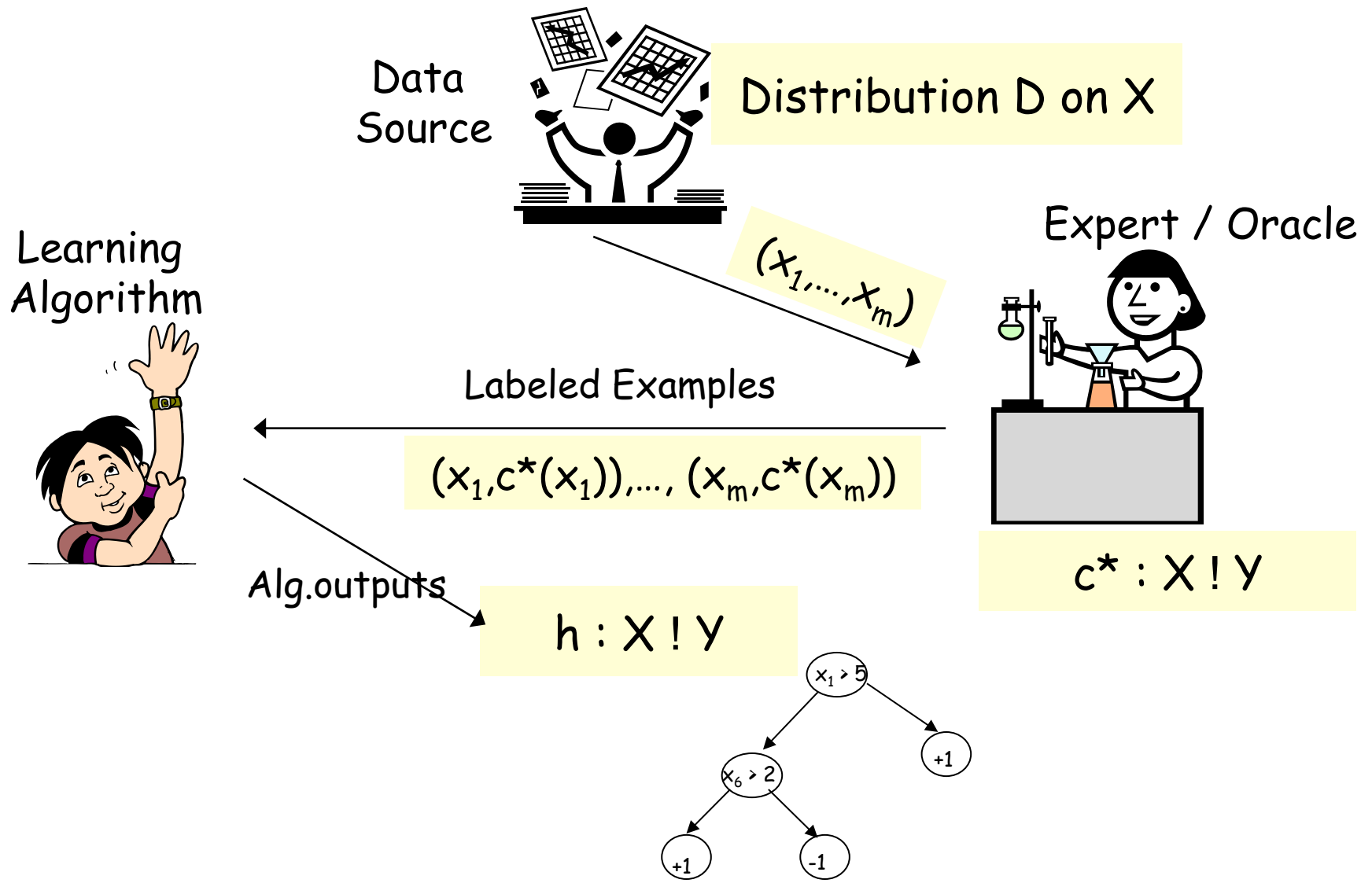
---



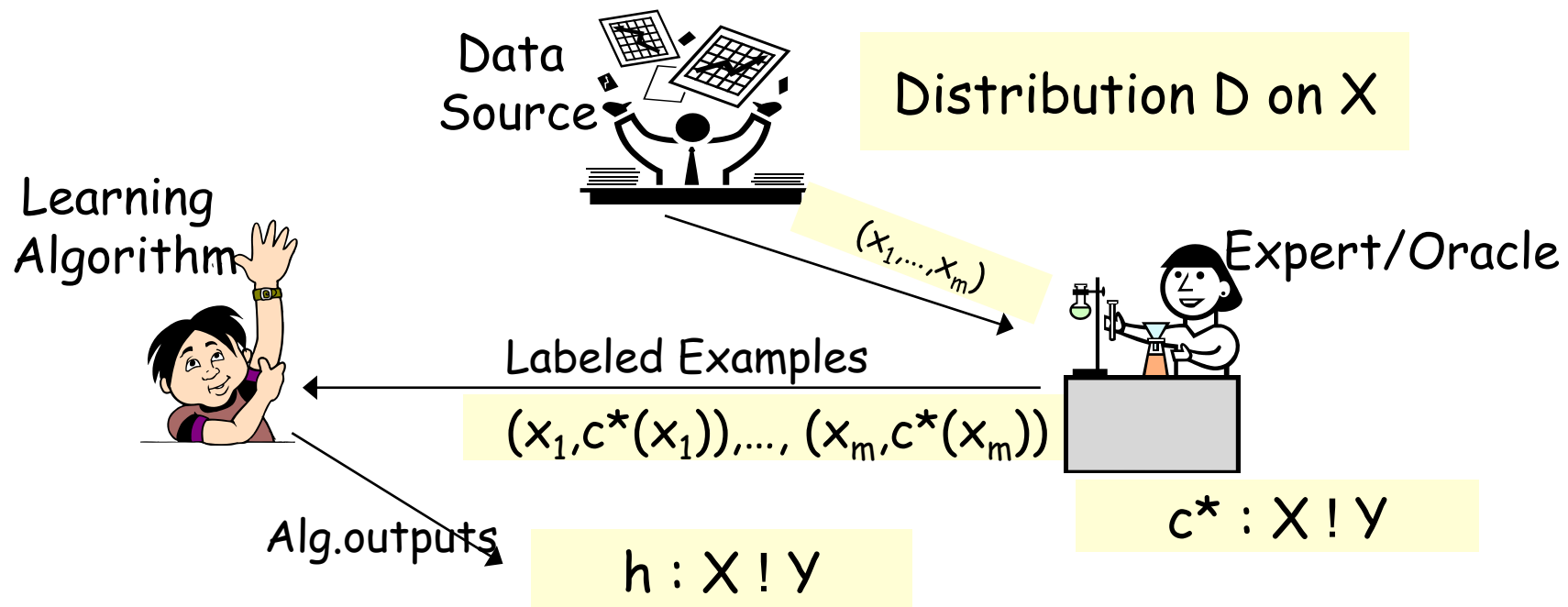


# Decision Tree Learning, Formal Guarantees

# Supervised Learning or Function Approximation



# Supervised Learning or Function Approximation



- Algo sees training sample  $S: (x_1, c^*(x_1)), \dots, (x_m, c^*(x_m))$ ,  $x_i$  i.i.d. from  $D$
- Does **optimization over  $S$** , finds hypothesis  $h$  (e.g., a decision tree).
- Goal:  $h$  has small error over  $D$ .

$$\text{err}(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$$

# Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

Confidence Bounds, Generalization



(Labeled) Data

Confidence for rule effectiveness on future data.

- Very well understood: Occam's bound, VC theory, etc.
- Decision trees: if we were able to find a small decision tree that explains data well, then good generalization guarantees.
  - NP-hard [Hyafil-Rivest'76]

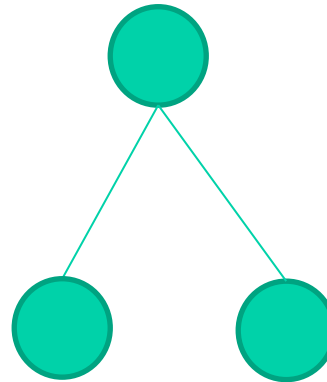


# Top Down Decision Trees Algorithms

- Decision trees: if we were able to find a **small decision tree** consistent with the data, then good generalization guarantees.
  - NP-hard [Hyafil-Rivest'76] 
- Very nice practical heuristics; top down algorithms, e.g, ID3
- Natural greedy approaches where we grow the tree from the root to the leaves by repeatedly replacing an existing leaf with an internal node.
  - Key point: **splitting criterion**.
  - ID3: split the leaf that decreases the entropy the most.
-  • Why not split according to error rate --- this is what we care about after all?
  - There are examples where we can get stuck in local minima!!!

# Entropy as a better splitting measure

|   |   |   |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |



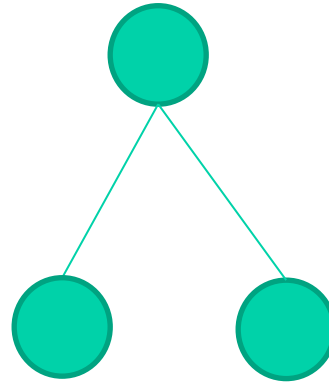
Initial error rate is  $1/4$  (25% positive, 75% negative)

Error rate after split is (left leaf is 100% negative;  
right leaf is 50/50)

Overall error doesn't decrease!

# Entropy as a better splitting measure

|   |   |   |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |




Initial entropy is

Entropy after split is

Entropy decreases!

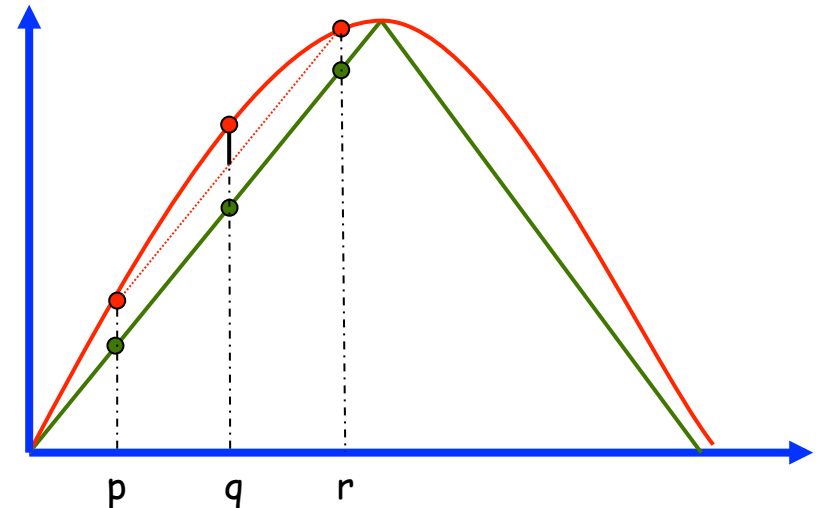
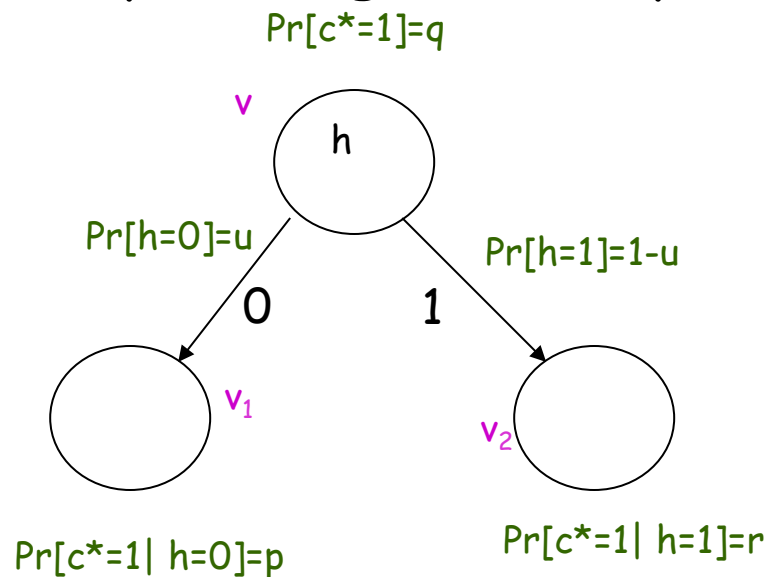
# Top Down Decision Trees Algorithms

- Natural greedy approaches where we grow the tree from the root to the leaves by repeatedly replacing an existing leaf with an internal node.
  - Key point: **splitting criterion**.
  - ID3: split the leaf that decreases the entropy the most.
-  Why not split according to error rate --- this is what we care about after all?
  - There are examples where you can get stuck!!!
- [Kearns-Mansour'96]: if measure of progress is entropy, we can always guarantee success under some formal relationships between the class of splits and the target (the class of splits can weakly approximate the target function).
  - Provides a way to think about the effectiveness of various top down algos.



# Top Down Decision Trees Algorithms

- Key: strong concavity of the splitting criterion



- $q = up + (1-u)r$ . Want to lower bound:  $G(q) - [uG(p) + (1-u)G(r)]$
- If:  $G(q) = \min(q, 1-q)$  (error rate), then  $G(q) = uG(p) + (1-u)G(r)$
- If:  $G(q) = H(q)$  (entropy), then  $G(q) - [uG(p) + (1-u)G(r)] > 0$  if  $r-p > 0$  and  $u \neq 1, u \neq 0$  (this happens under the weak learning assumption)

# Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.

# What you should know:

---

- Well posed function approximation problems:
  - Instance space,  $X$
  - Sample of labeled training data  $\{ \langle x^{(i)}, y^{(i)} \rangle \}$
  - Hypothesis space,  $H = \{ f: X \rightarrow Y \}$
- 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 (?)
  - But inductive learning without some bias is futile !
- Decision tree learning
  - Greedy top-down learning of decision trees (ID3, C4.5, ...)
  - Overfitting and tree post-pruning
  - Extensions...

# Extra slides

extensions to decision tree learning

# Continuous Valued Attributes

---

Create a discrete attribute to test continuous

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

|                     |    |    |     |     |     |    |
|---------------------|----|----|-----|-----|-----|----|
| <i>Temperature:</i> | 40 | 48 | 60  | 72  | 80  | 90 |
| <i>PlayTennis:</i>  | No | No | Yes | Yes | Yes | No |

# Rule Post-Pruning

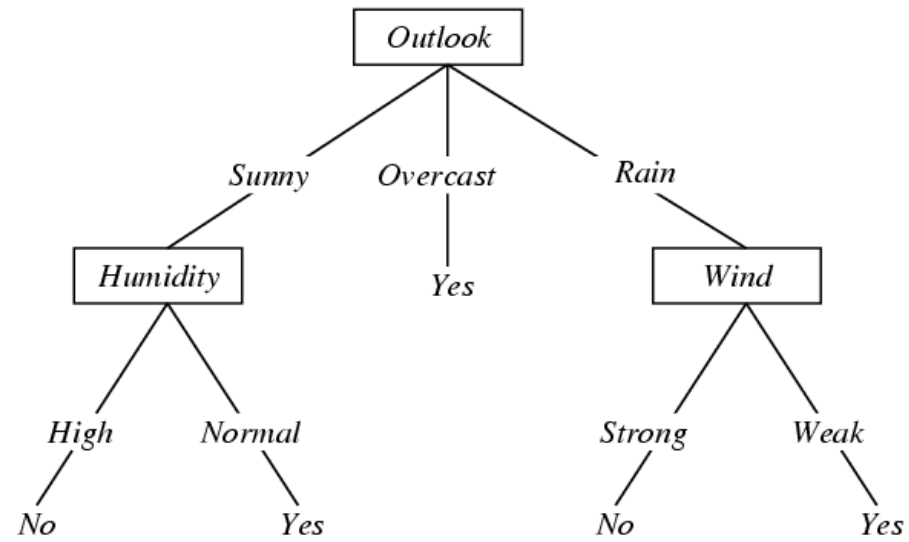
---

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

frequently used method (e.g., C4.5)

# Converting A Tree to Rules

---



# 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

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

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

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



# Unknown Attribute Values

---

What if some examples missing values of  $A$ ?

Use training example anyway, sort through tree

- If node  $n$  tests  $A$ , assign most common value of  $A$  among other examples sorted to node  $n$
- assign most common value of  $A$  among other examples with same target value
- assign probability  $p_i$  to each possible value  $v_i$  of  $A$ 
  - assign fraction  $p_i$  of example to each descendant in tree

Classify new examples in same fashion

## 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  $f: \langle x_1, x_2 \rangle \rightarrow y$ , where  $x_1$  and  $x_2$  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:  $12/20 = .60$



# Machine Learning 10-601

Tom M. Mitchell  
Machine Learning Department  
Carnegie Mellon University

January 14, 2015

Today:

- Review: probability

many of these slides are  
derived from William Cohen,  
Andrew Moore, Aarti Singh,  
Eric Xing. Thanks!

Readings:

Probability review

- Bishop Ch. 1 thru 1.2.3
- Bishop, Ch. 2 thru 2.2
- Andrew Moore's online tutorial

# Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence

# Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment
- Examples
  - A = True if a randomly drawn person from our class is female
  - A = The hometown of a randomly drawn person from our class
  - A = True if two randomly drawn persons from our class have same birthday
- Define  $P(A)$  as “the fraction of possible worlds in which A is true” or “the fraction of times A holds, in repeated runs of the random experiment”
  - the set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S
$$A: S \rightarrow \{0,1\}$$



# A little formalism

More formally, we have

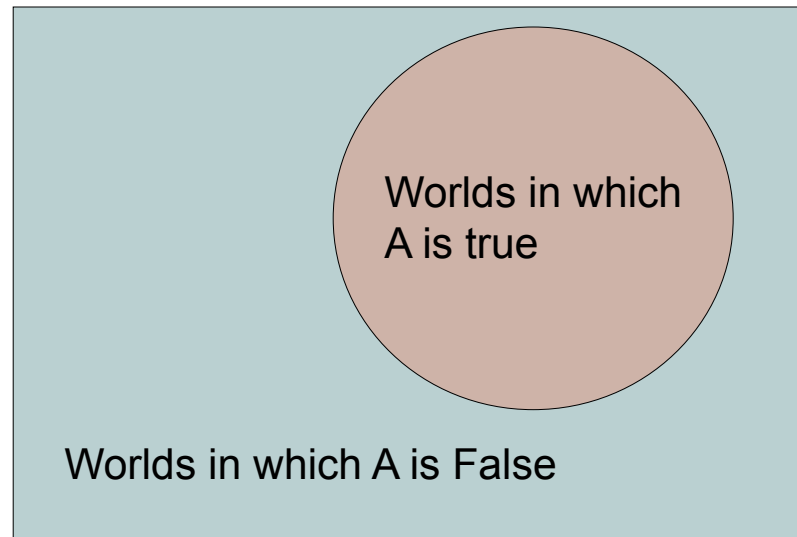
- a sample space  $S$  (e.g., set of students in our class)
  - aka the set of possible worlds
- a random variable is a function defined over the sample space
  - Gender:  $S \rightarrow \{m, f\}$
  - Height:  $S \rightarrow \text{Reals}$
- an event is a subset of  $S$ 
  - e.g., the subset of  $S$  for which Gender=f
  - e.g., the subset of  $S$  for which (Gender=m) AND (eyeColor=blue)
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

# Visualizing A

Sample space  
of all possible  
worlds



Its area is 1



$P(A)$  = Area of  
reddish oval

# The Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

[di Finetti 1931]:

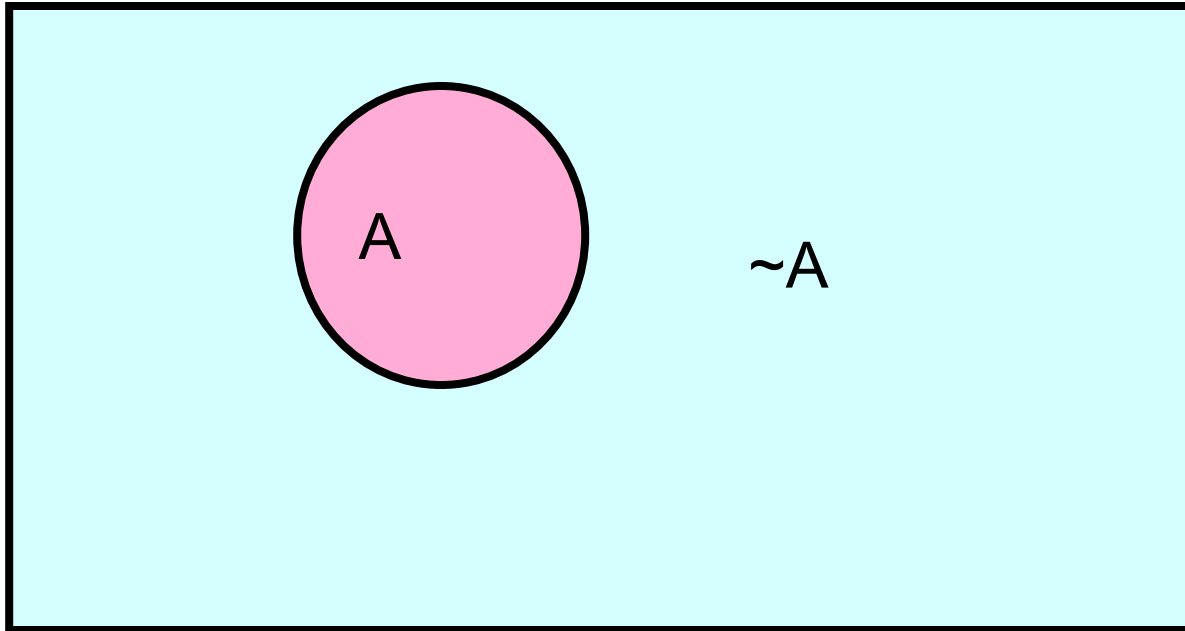
when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

# Elementary Probability in Pictures

- $P(\sim A) + P(A) = 1$



# A useful theorem

- $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$ ,  
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$$\rightarrow P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

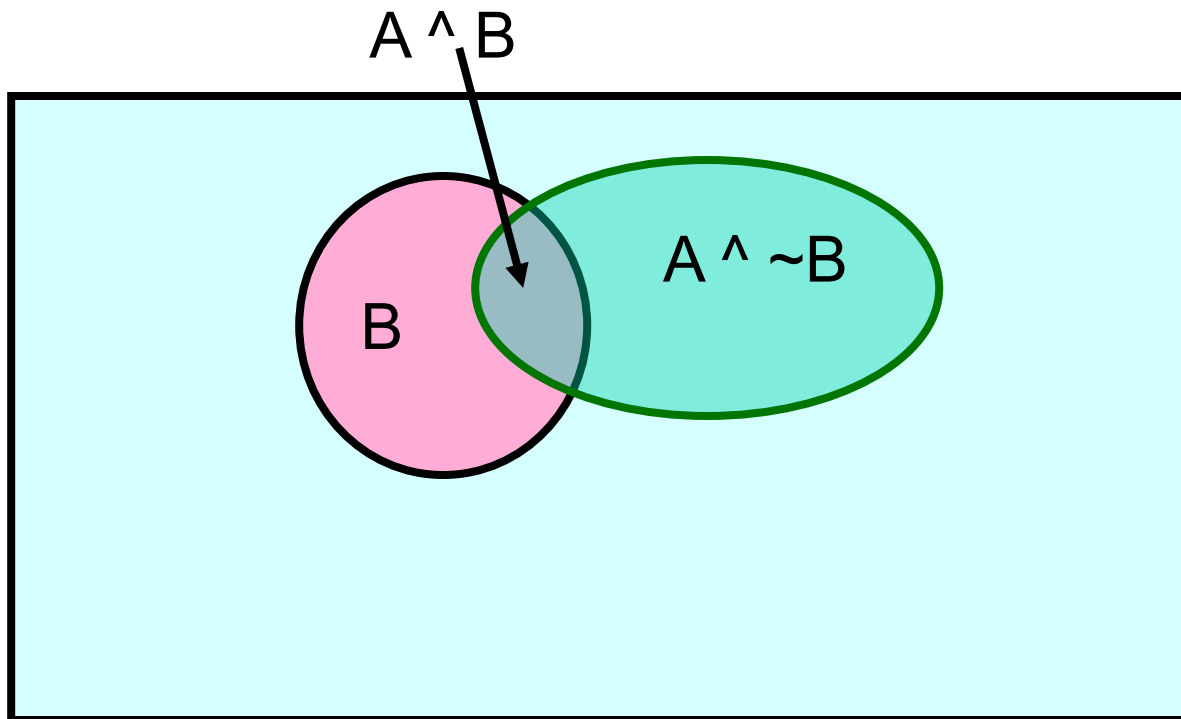
$$A = [A \text{ and } (B \text{ or } \sim B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)]$$

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P((A \text{ and } B) \text{ and } (A \text{ and } \sim B))$$

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - \cancel{P(A \text{ and } B \text{ and } A \text{ and } \sim B)}$$

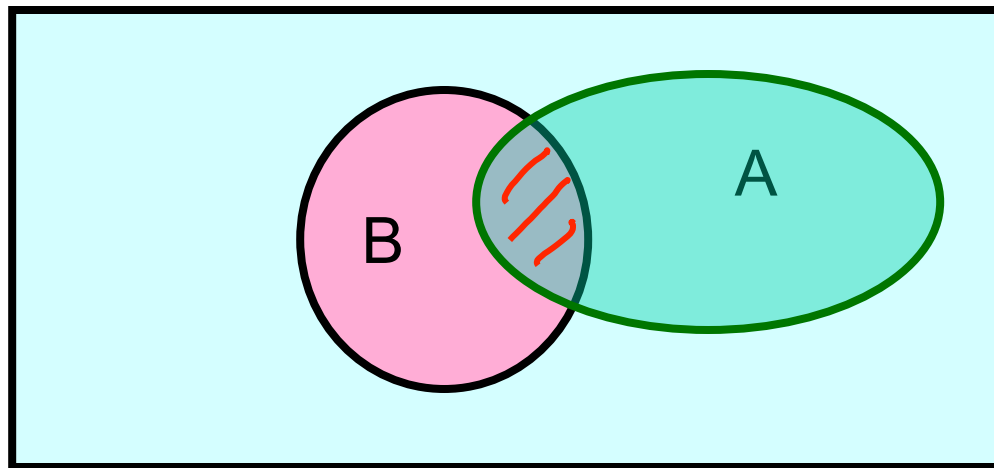
# Elementary Probability in Pictures

- $P(A) = P(A \wedge B) + P(A \wedge \sim B)$



# Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$



# Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

## Corollary: The Chain Rule

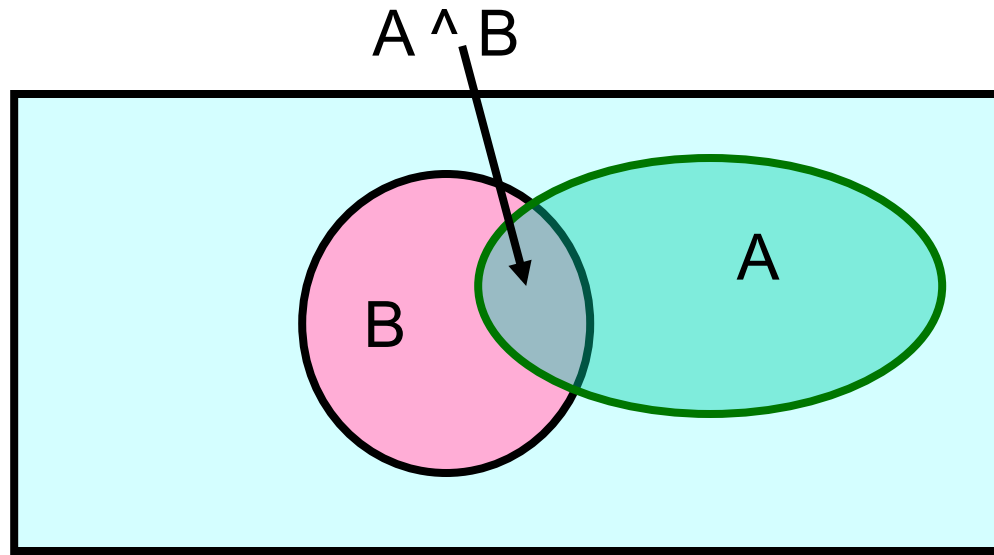
$$\underline{P(A \wedge B)} = P(A|B) P(B)$$

$$\begin{aligned} P(A \wedge B \wedge C) &= P(A|B \wedge C) \underbrace{P(B \wedge C)} \\ &= P(A|B \wedge C) P(B|C) P(C) \end{aligned}$$



# Bayes Rule

- let's write 2 expressions for  $P(A \wedge B)$



$$\begin{aligned} P(A \wedge B) &= P(A|B) P(B) \\ &= P(B|A) P(A) \end{aligned}$$
$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad \text{Bayes' rule}$$

we call  $P(A)$  the “prior”

and  $P(A|B)$  the “posterior”



**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

...by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter.... necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*...

## Other Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

# Applying Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

A = you have the flu, B = you just coughed

Assume:

$$P(A) = 0.05$$

$$P(B|A) = 0.80$$

$$P(B|\sim A) = 0.2$$

what is  $P(\text{flu} | \text{cough}) = P(A|B)$ ?

what does all this have to do with  
function approximation?

# The Joint Distribution

*Example: Boolean  
variables  $A, B, C$*

Recipe for making a joint  
distribution of  $M$  variables:

# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).

| <b>A</b> | <b>B</b> | <b>C</b> |
|----------|----------|----------|
| 0        | 0        | 0        |
| 0        | 0        | 1        |
| 0        | 1        | 0        |
| 0        | 1        | 1        |
| 1        | 0        | 0        |
| 1        | 0        | 1        |
| 1        | 1        | 0        |
| 1        | 1        | 1        |

# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.

| <b>A</b> | <b>B</b> | <b>C</b> | <b>Prob</b> |
|----------|----------|----------|-------------|
| 0        | 0        | 0        | 0.30        |
| 0        | 0        | 1        | 0.05        |
| 0        | 1        | 0        | 0.10        |
| 0        | 1        | 1        | 0.05        |
| 1        | 0        | 0        | 0.05        |
| 1        | 0        | 1        | 0.10        |
| 1        | 1        | 0        | 0.25        |
| 1        | 1        | 1        | 0.10        |

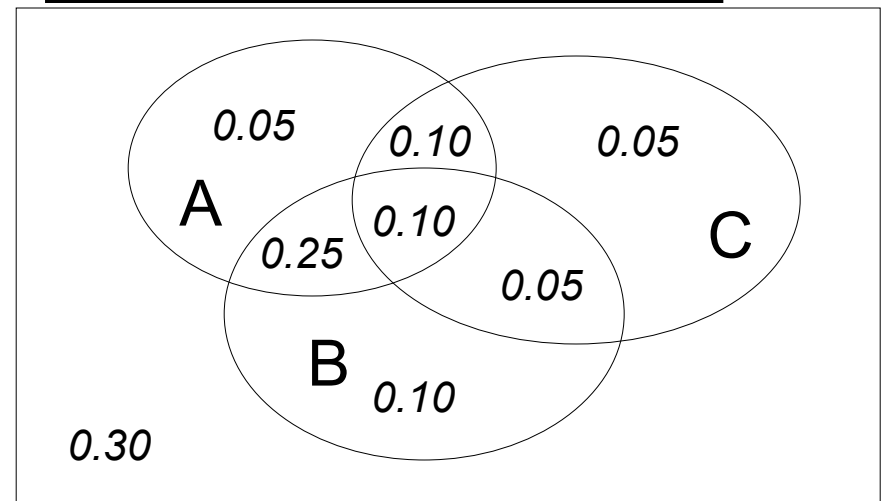


# The Joint Distribution









Recipe for making a joint distribution of  $M$  variables:

1. Make a truth table listing all combinations of values of your variables (if there are  $M$  Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

| <b>A</b> | <b>B</b> | <b>C</b> | <b>Prob</b> |
|----------|----------|----------|-------------|
| 0        | 0        | 0        | 0.30        |
| 0        | 0        | 1        | 0.05        |
| 0        | 1        | 0        | 0.10        |
| 0        | 1        | 1        | 0.05        |
| 1        | 0        | 0        | 0.05        |
| 1        | 0        | 1        | 0.10        |
| 1        | 1        | 0        | 0.25        |
| 1        | 1        | 1        | 0.10        |











# Using the Joint Distribution

| gender | hours_worked | wealth |           |   |
|--------|--------------|--------|-----------|---|
| Female | v0:40.5-     | poor   | 0.253122  |  |
|        |              | rich   | 0.0245895 |  |
|        | v1:40.5+     | poor   | 0.0421768 |  |
|        |              | rich   | 0.0116293 |  |
| Male   | v0:40.5-     | poor   | 0.331313  |  |
|        |              | rich   | 0.0971295 |  |
|        | v1:40.5+     | poor   | 0.134106  |  |
|        |              | rich   | 0.105933  |  |

Once you have the JD  
you can ask for the  
probability of any logical  
expression involving  
your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Using the Joint

| gender | hours_worked | wealth |   |
|--------|--------------|--------|---|
| Female | v0:40.5-     | poor   | 0.253122   |
|        |              | rich   | 0.0245895  |
|        | v1:40.5+     | poor   | 0.0421768  |
|        |              | rich   | 0.0116293  |
| Male   | v0:40.5-     | poor   | 0.331313   |
|        |              | rich   | 0.0971295  |
|        | v1:40.5+     | poor   | 0.134106   |
|        |              | rich   | 0.105933   |

$$P(\text{Poor Male}) = 0.4654$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Using the Joint

| gender | hours_worked | wealth |           |
|--------|--------------|--------|-----------|
| Female | v0:40.5-     | poor   | 0.253122  |
|        |              | rich   | 0.0245895 |
|        | v1:40.5+     | poor   | 0.0421768 |
|        |              | rich   | 0.0116293 |
| Male   | v0:40.5-     | poor   | 0.331313  |
|        |              | rich   | 0.0971295 |
|        | v1:40.5+     | poor   | 0.134106  |
|        |              | rich   | 0.105933  |

$$P(\text{Poor}) = 0.7604$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Inference with the Joint

| gender | hours_worked | wealth |           |             |
|--------|--------------|--------|-----------|-------------|
| Female | v0:40.5-     | poor   | 0.253122  | <div></div> |
|        |              | rich   | 0.0245895 | <div></div> |
|        | v1:40.5+     | poor   | 0.0421768 | <div></div> |
|        |              | rich   | 0.0116293 | <div></div> |
| Male   | v0:40.5-     | poor   | 0.331313  | <div></div> |
|        |              | rich   | 0.0971295 | <div></div> |
|        | v1:40.5+     | poor   | 0.134106  | <div></div> |
|        |              | rich   | 0.105933  | <div></div> |

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

# Inference with the Joint

| gender | hours_worked | wealth |           |
|--------|--------------|--------|-----------|
| Female | v0:40.5-     | poor   | 0.253122  |
|        |              | rich   | 0.0245895 |
|        | v1:40.5+     | poor   | 0.0421768 |
|        |              | rich   | 0.0116293 |
| Male   | v0:40.5-     | poor   | 0.331313  |
|        |              | rich   | 0.0971295 |
|        | v1:40.5+     | poor   | 0.134106  |
|        |              | rich   | 0.105933  |

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$

# You should know

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Conditional probabilities
- Chain rule
- Bayes rule
- Joint distribution over multiple random variables
  - how to calculate other quantities from the joint distribution

# Expected values

Given discrete random variable  $X$ , the expected value of  $X$ , written  $E[X]$  is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

We also can talk about the expected value of functions of  $X$

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P(X = x)$$



# Covariance

Given two discrete r.v.'s  $X$  and  $Y$ , we define the covariance of  $X$  and  $Y$  as

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

e.g.,  $X=\text{gender}$ ,  $Y=\text{playsFootball}$

or  $X=\text{gender}$ ,  $Y=\text{leftHanded}$

Remember:

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$