



Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 15, 2011

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

Top-Down Induction of Decision Trees

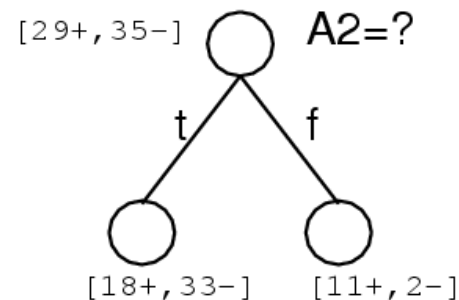
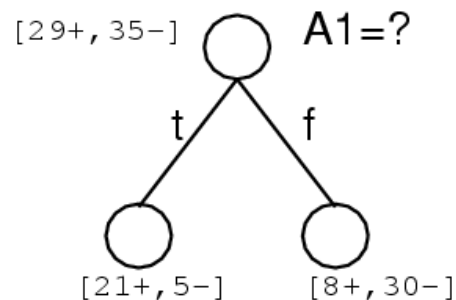
[ID3, C4.5, Quinlan]

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

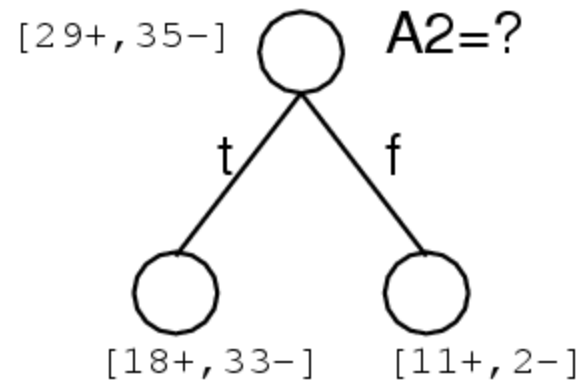
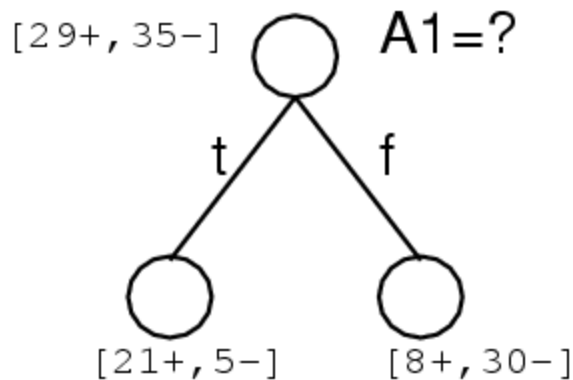
Which attribute is best?



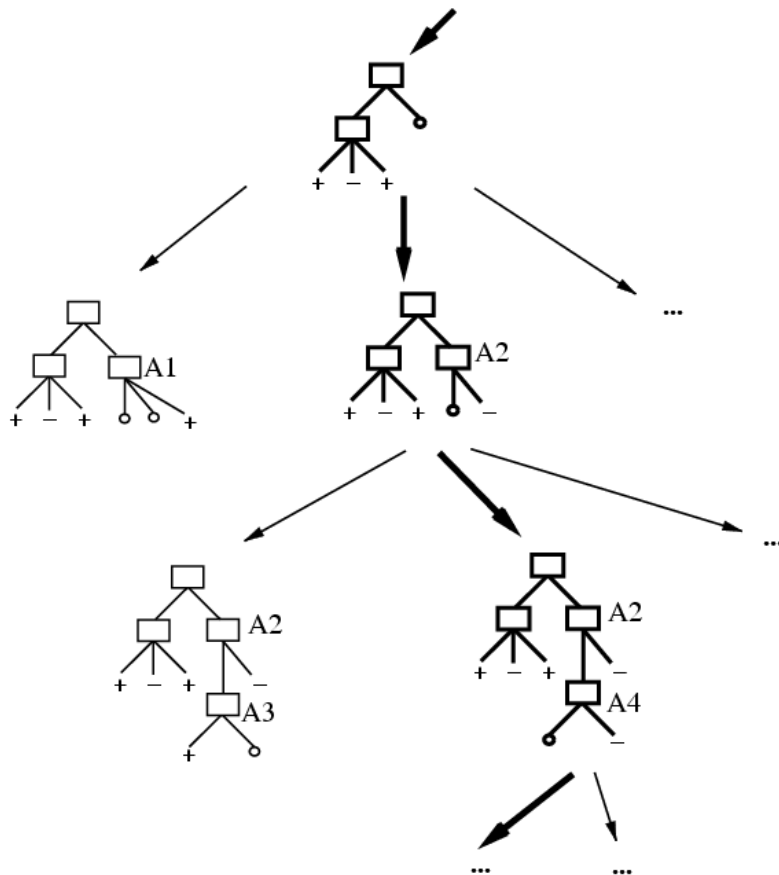
Information Gain (also called mutual information)
between input attribute A and target variable Y

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

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



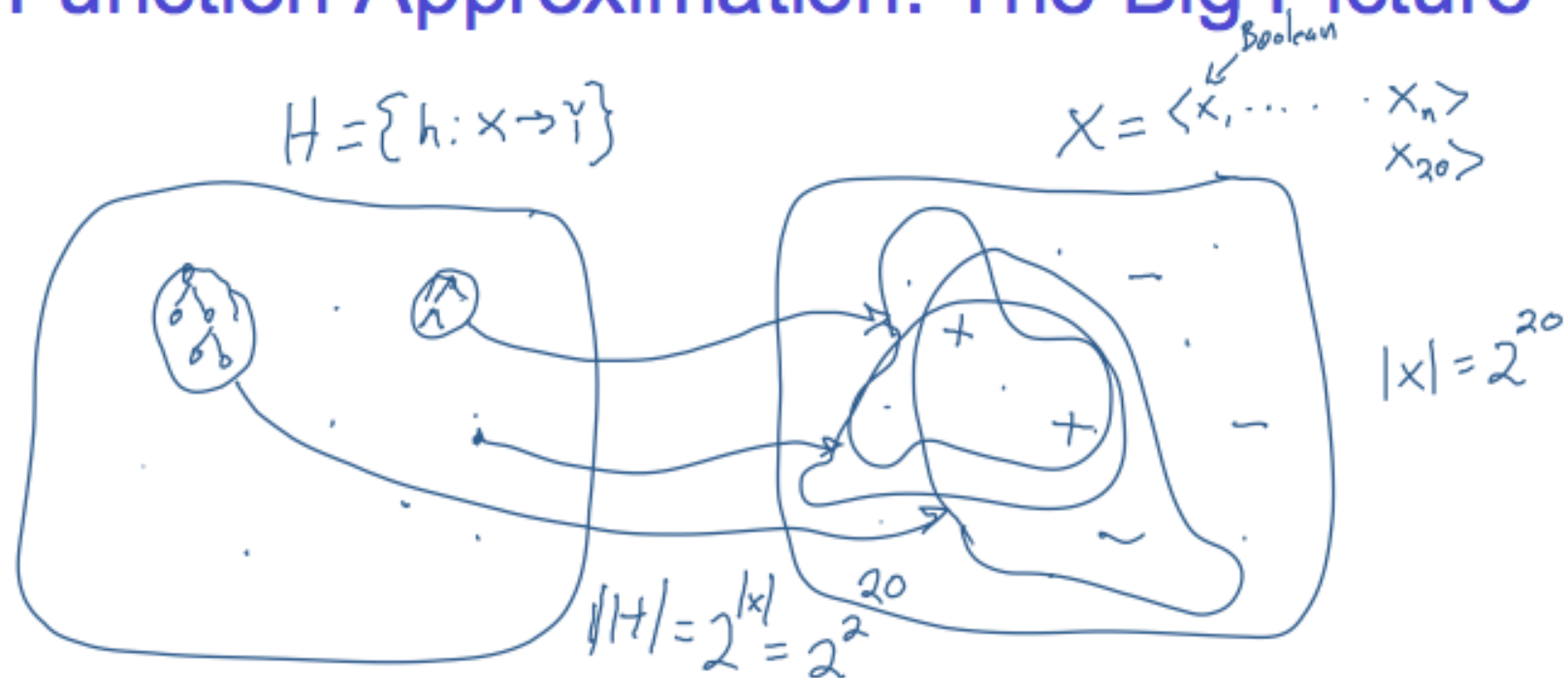
Function approximation as Search for the best hypothesis



- ID3 performs heuristic search through space of decision trees

Function Approximation: The Big Picture

Function Approximation: The Big Picture



How many labeled examples are needed in order to determine which of the 2^{20} hypotheses is the correct one?

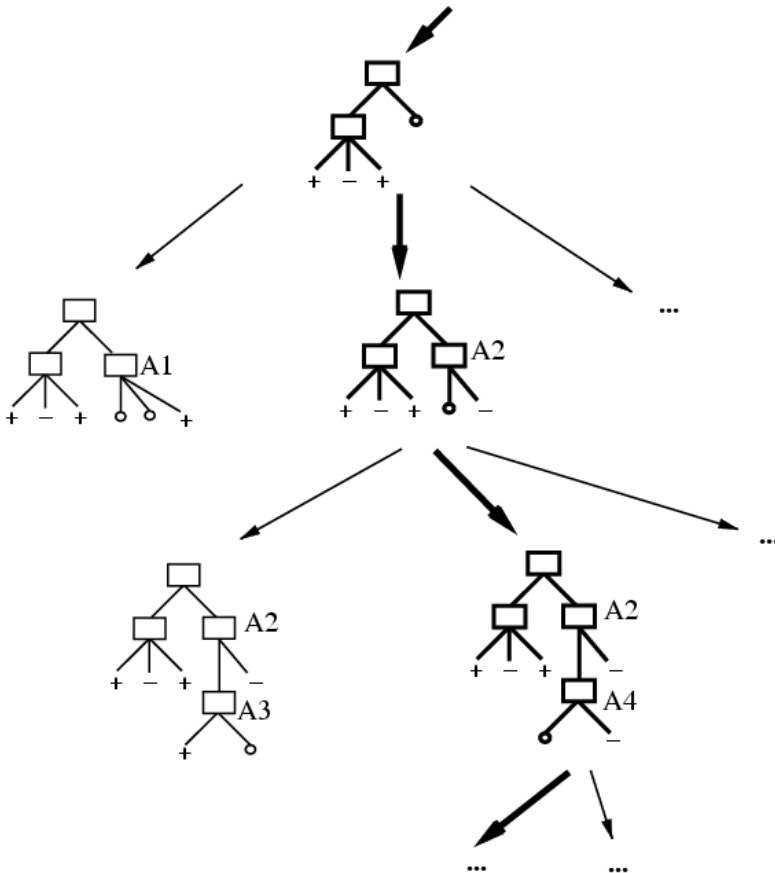
All 2^{20} instances in X must be labeled!

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g. priors over H)

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:

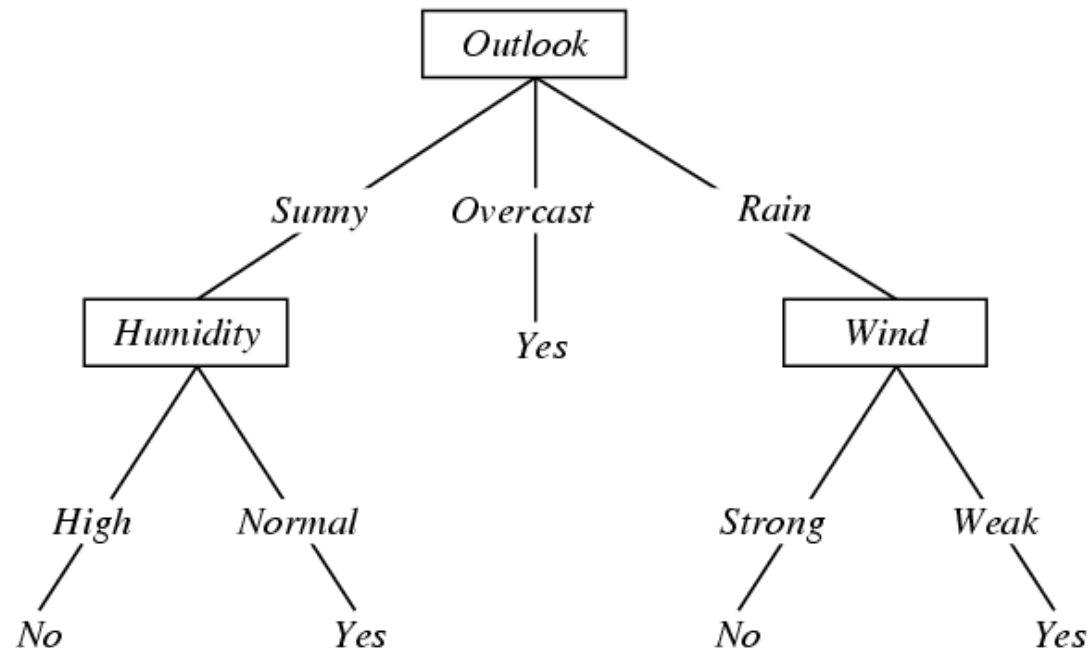
- Also fewer hypotheses containing a prime number of nodes and attributes beginning with “Z”
- What's so special about “short” hypotheses, instead of “prime number of nodes”?

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: $error_{true}(h)$

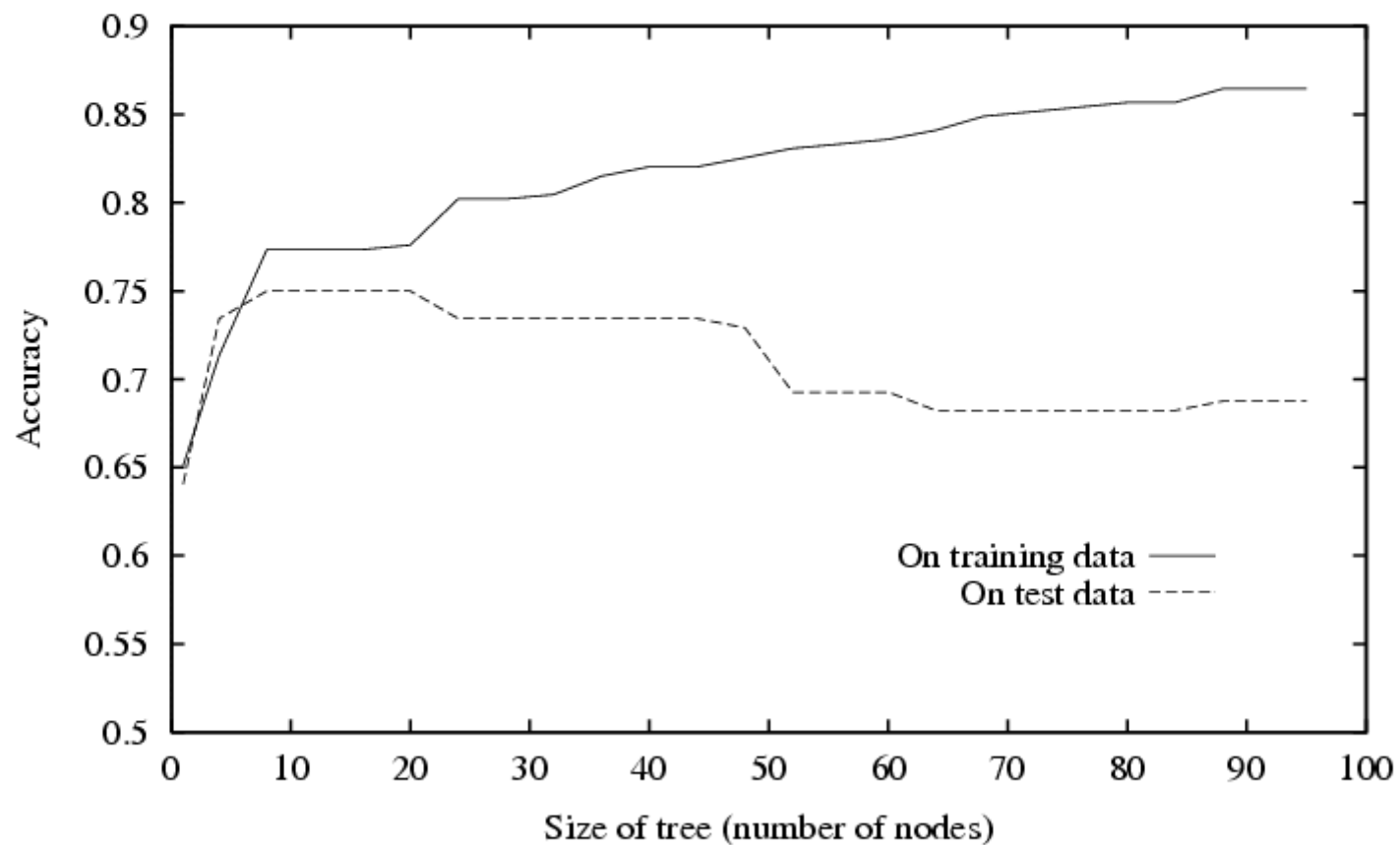
We say h overfits the training data if

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

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

Avoiding Overfitting

How can we avoid overfitting?

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

How to select “best” tree:

- Measure performance over training data
- Measure performance over separate validation data set
- MDL: minimize
 $size(tree) + size(misclassifications(tree))$

Reduced-Error Pruning

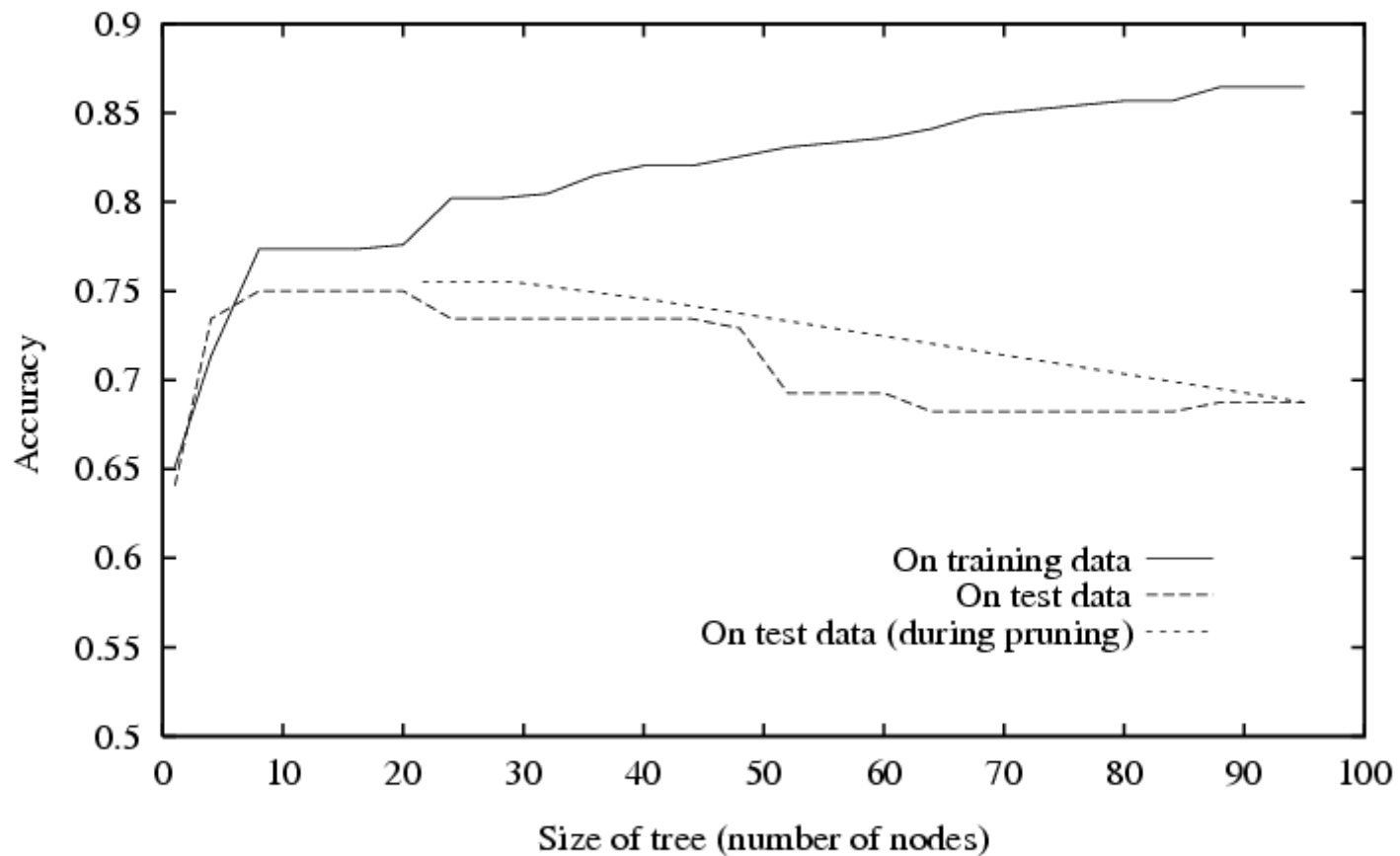
Split data into *training* and *validation* set

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

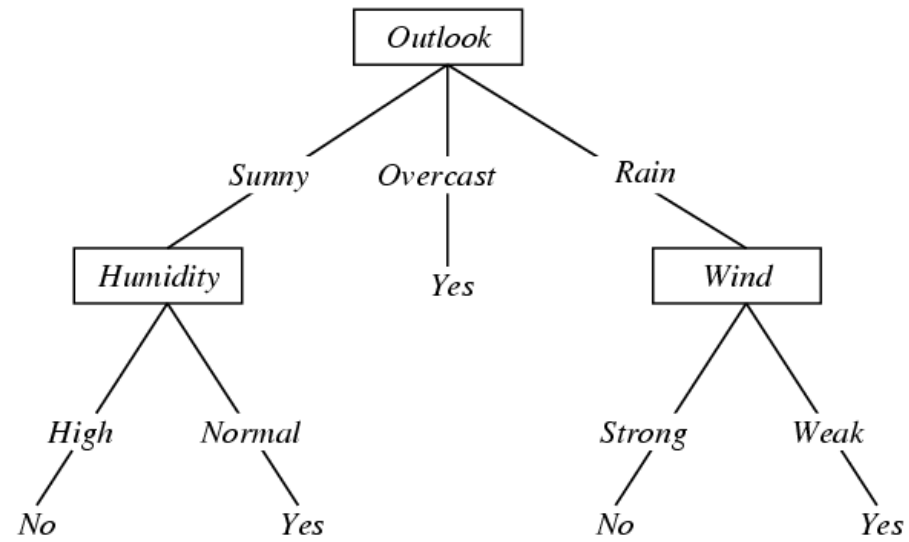


Rule Post-Pruning

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

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

Converting A Tree to Rules



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

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

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?



Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 15, 2011

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

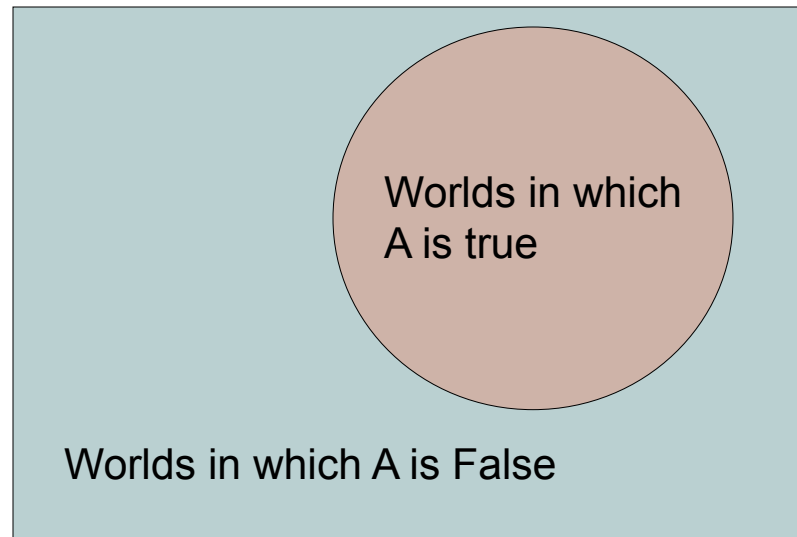
- a sample space S (e.g., set of students in our class)
 - aka the set of possible worlds
- a random variable is a function defined over the sample space
 - Gender: $S \rightarrow \{m, f\}$
 - Height: $S \rightarrow \text{Reals}$
- an event is a subset of S
 - e.g., the subset of S for which Gender=f
 - 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]:

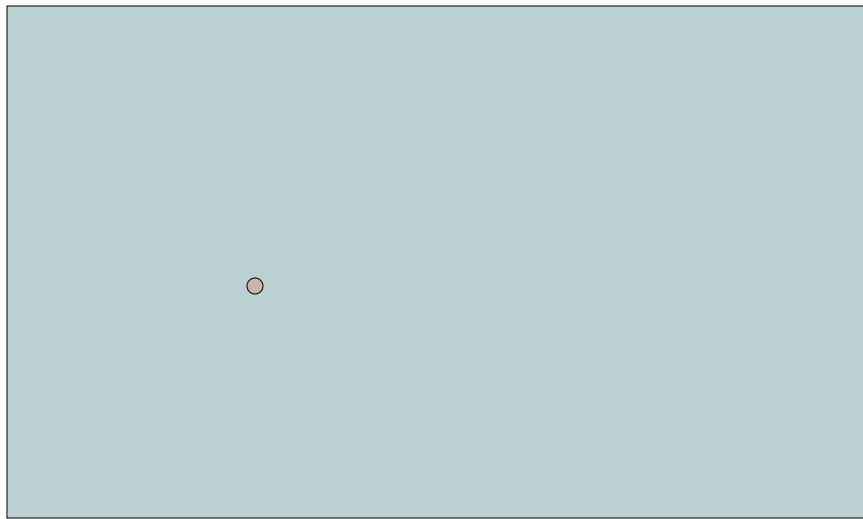
when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

Interpreting the axioms

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

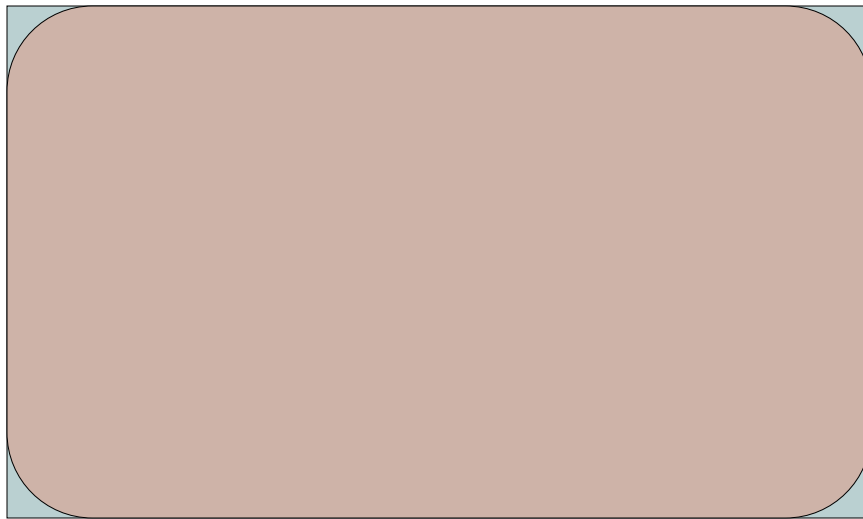


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

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



The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

Interpreting the axioms

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

Theorems from the Axioms

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
 - ➔ $P(\text{not } A) = P(\sim A) = 1 - P(A)$

Theorems from the Axioms

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
 $\rightarrow P(\text{not } A) = P(\sim A) = 1 - P(A)$

$$P(A \text{ or } \sim A) = 1$$

$$P(A \text{ and } \sim A) = 0$$

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



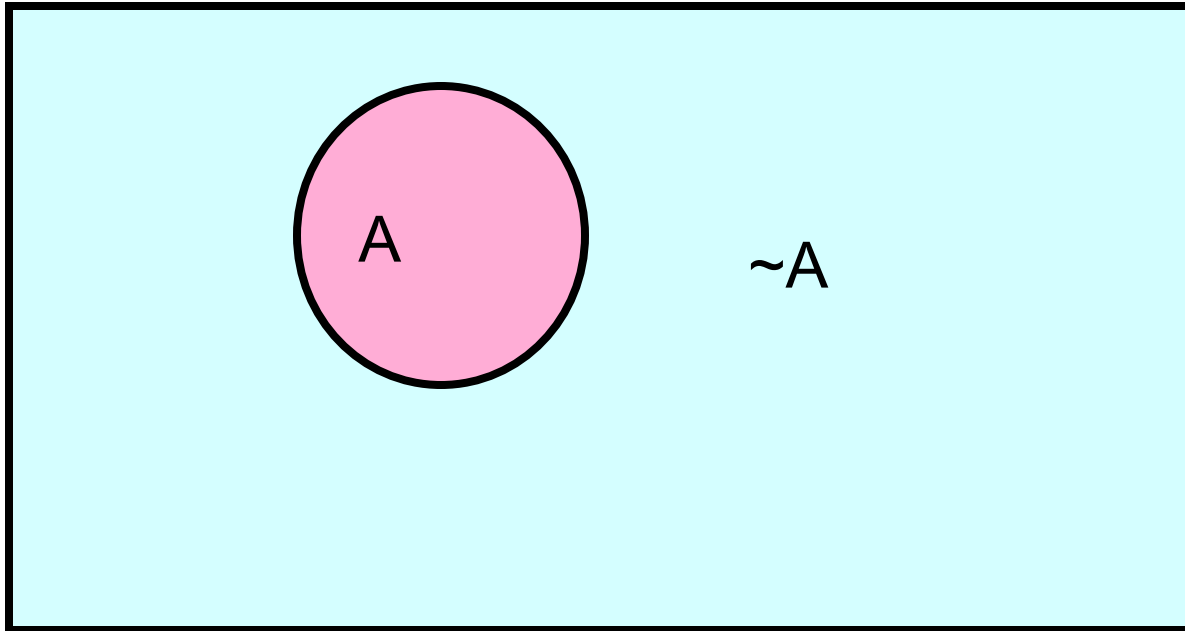
1



$$= P(A) + P(\sim A) + 0$$

Elementary Probability in Pictures

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



Another useful theorem

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

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

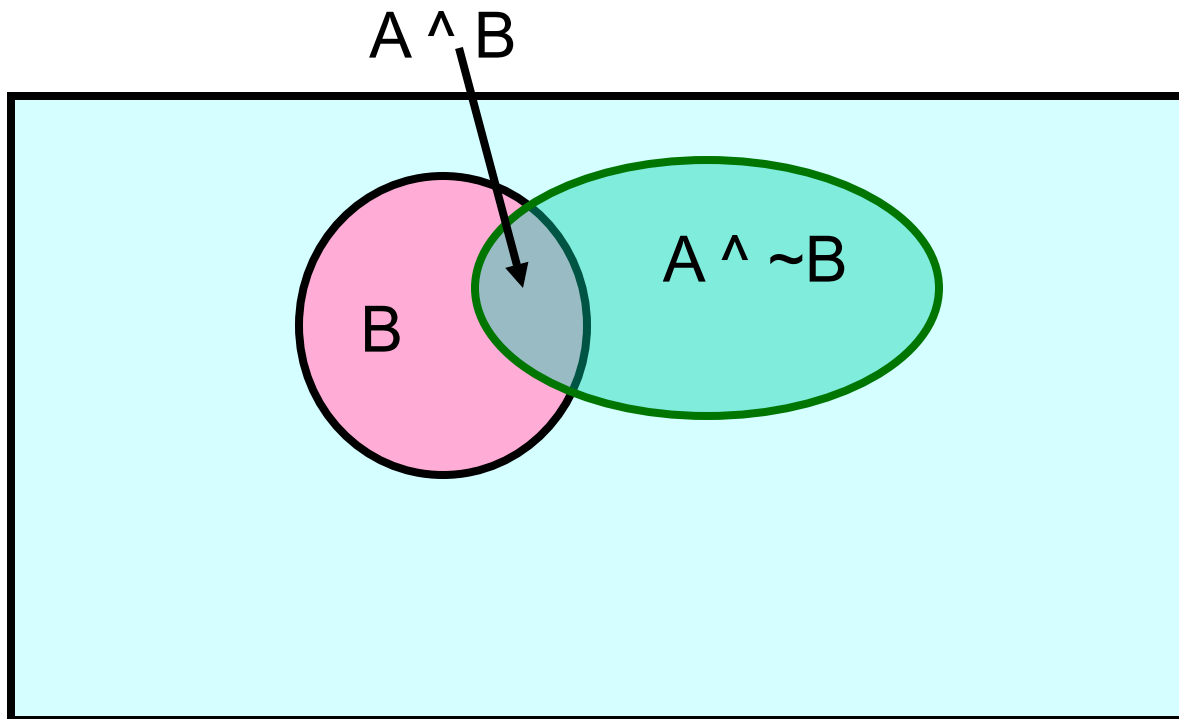
$$A = [A \text{ and } (B \text{ or } \sim B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)]$$

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

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

Elementary Probability in Pictures

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



Multivalued Discrete Random Variables

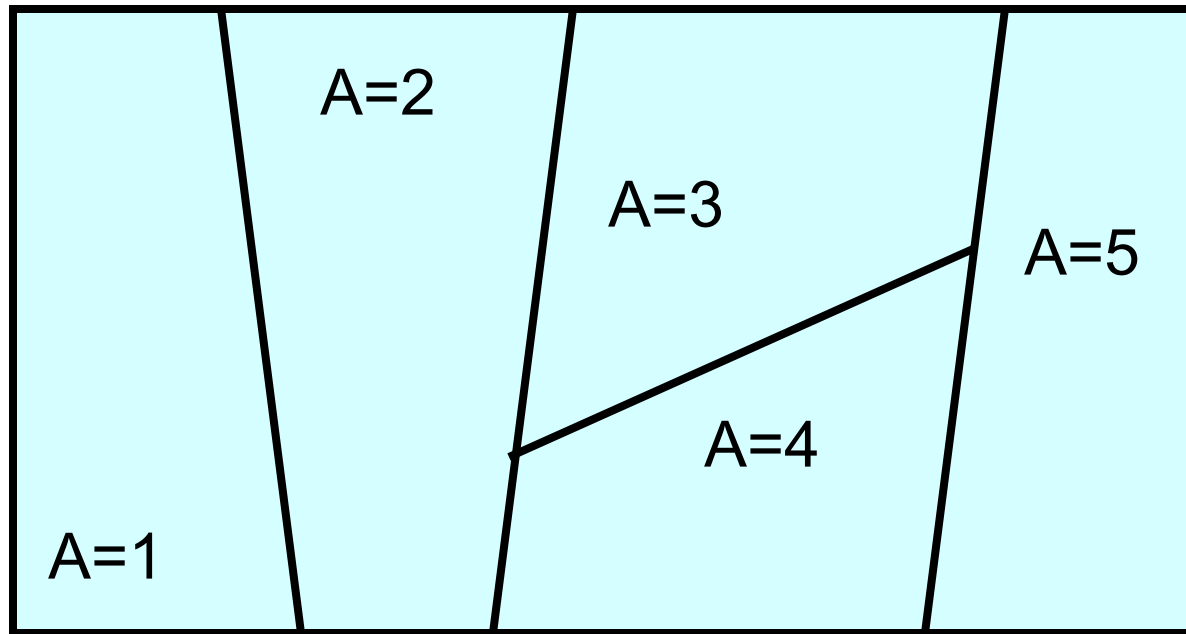
- Suppose A can take on more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \dots, v_k\}$

- Thus... $P(A = v_i \wedge A = v_j) = 0$ if $i \neq j$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

Elementary Probability in Pictures

$$\sum_{j=1}^k P(A = v_j) = 1$$



Definition of Conditional Probability

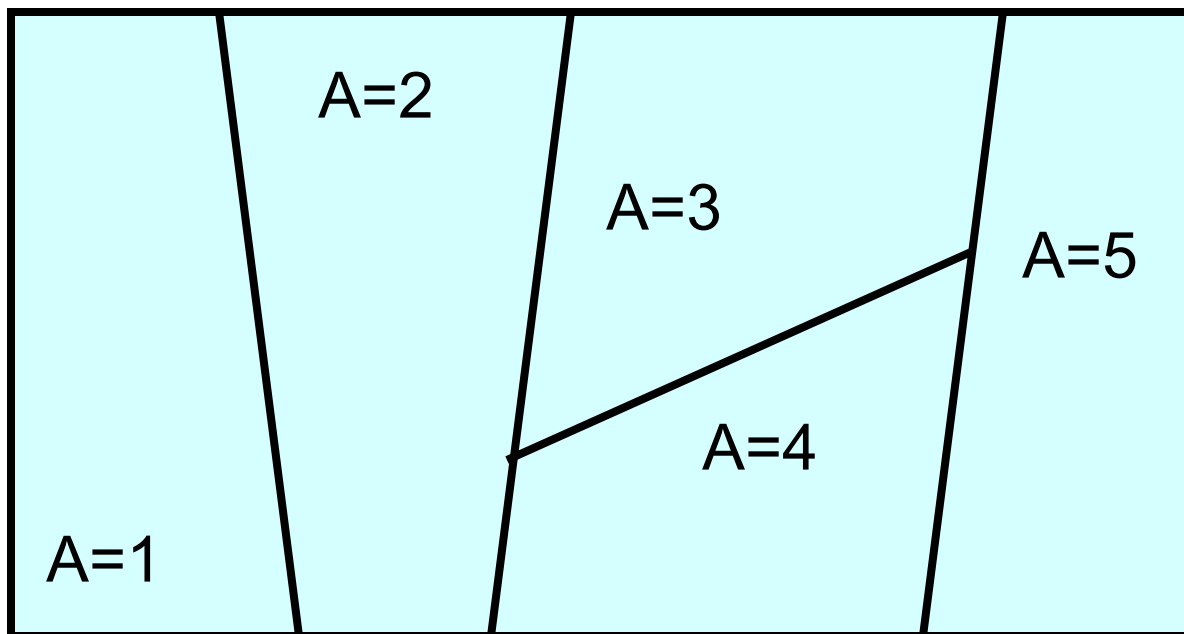
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B) P(B)$$

Conditional Probability in Pictures

picture: $P(B|A=2)$



Independent Events

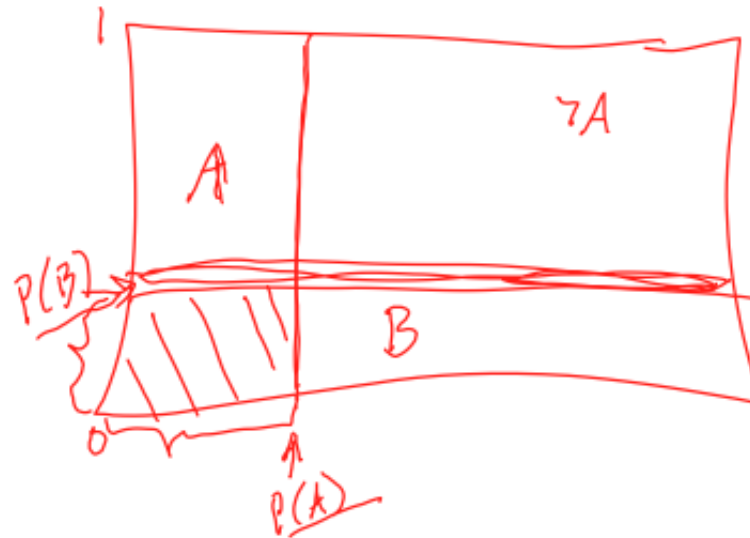
- Definition: two events A and B are *independent* if $\Pr(A \text{ and } B) = \Pr(A) * \Pr(B)$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Picture “A independent of B”

Independent Events

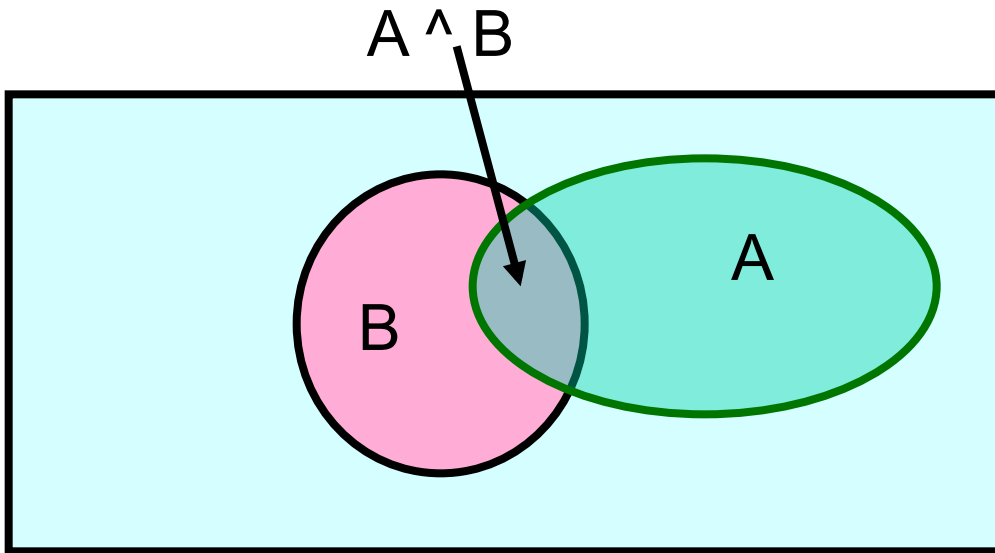
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Definition: two events A and B are *independent* if $P(A \cap B) = P(A) \cdot P(B)$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)



Elementary Probability in Pictures

- Let's write 2 expressions for $P(A \cap B)$



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

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

You should know

- 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

what does all this have to do with
function approximation?

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:

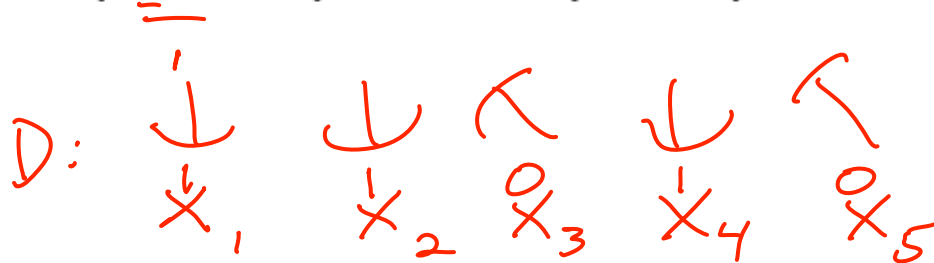
- ☐ He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
- ☐ You say: Please flip it a few times:



- ☐ You say: The probability is:
- ☐ **He says: Why???**
- ☐ You say: Because...

Thumbtack – Binomial Distribution

- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$



- Flips are i.i.d.:
 - Independent events
 - Identically distributed according to Binomial distribution
- Sequence \mathcal{D} of α_H Heads and α_T Tails


$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
 - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Maximum Likelihood Estimate for Θ


$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$

■ Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$$




■ Set derivative to zero:

$$\hat{\theta} = \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta)$$

$$= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$$

How many flips do I need?


$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

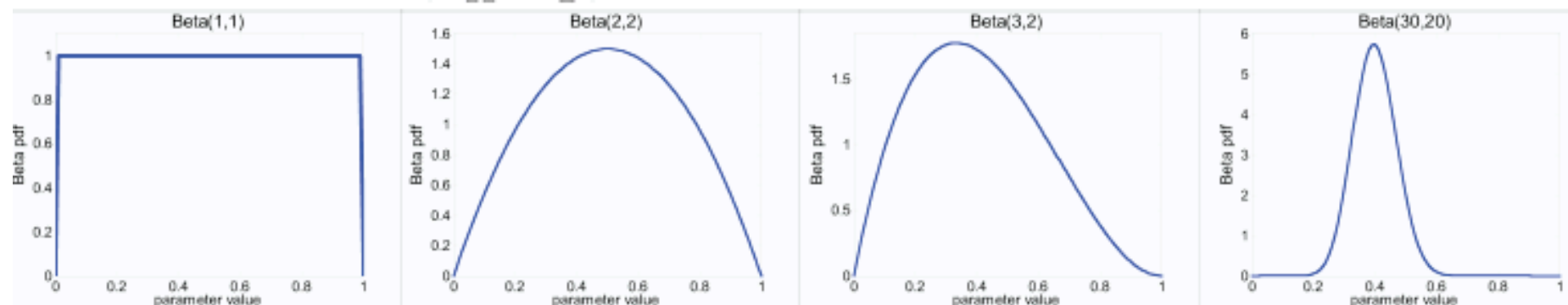
$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Mean:

Mode:



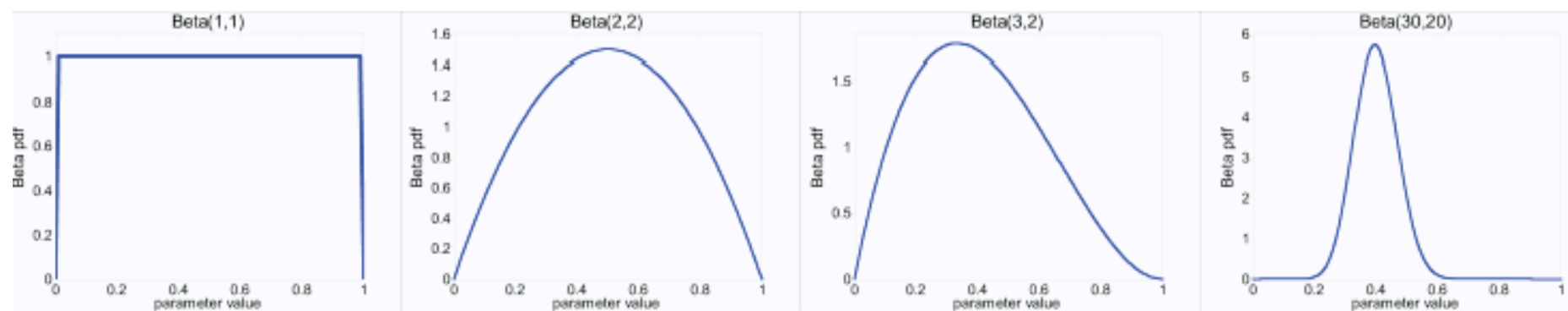
- Likelihood function: $P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

Posterior distribution

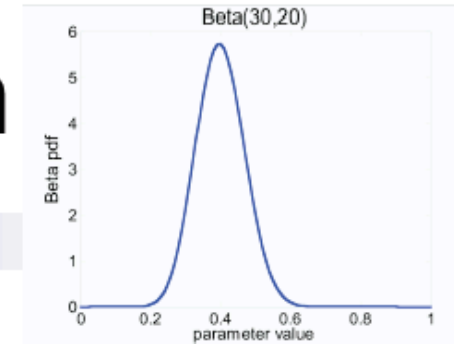
- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

Dirichlet distribution



- number of heads in N flips of a two-sided coin
 - follows a binomial distribution
 - Beta is a good prior (conjugate prior for binomial)
- what if it's not two-sided, but k-sided?
 - follows a multinomial distribution
 - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \dots, \theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_i - 1)}$$

Lejeune Dirichlet



Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	 Germany
Nationality	 German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data \mathcal{D}

$$\hat{\theta} = \arg \max_{\theta} P(\mathcal{D} | \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\theta | \mathcal{D}) \\ &= \arg \max_{\theta} = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}\end{aligned}$$

You should know

- Probability basics
 - random variables, events, sample space, conditional probs, ...
 - independence of random variables
 - Bayes rule
 - Joint probability distributions
 - calculating probabilities from the joint distribution
- Point estimation
 - maximum likelihood estimates
 - maximum a posteriori estimates
 - distributions – binomial, Beta, Dirichlet, ...

Extra slides

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).

A	B	C
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.

A	B	C	Prob
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

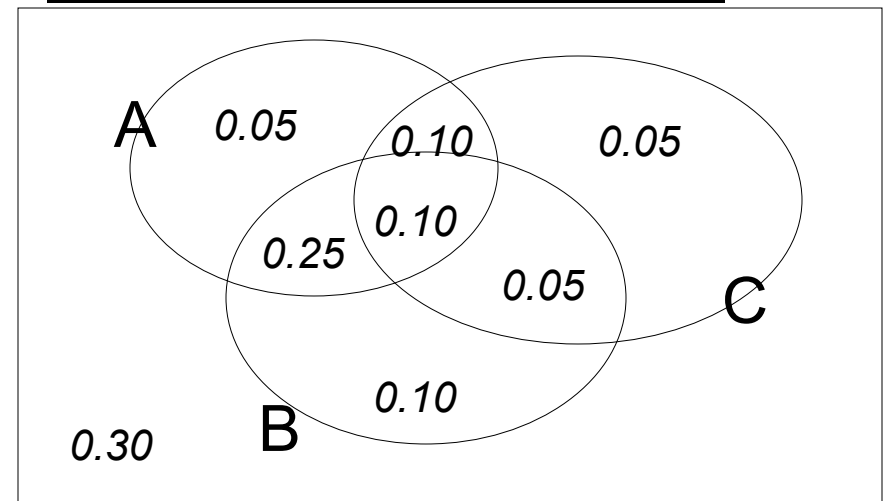
The Joint Distribution

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.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
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

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	<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>

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

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

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

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

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

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

Rememb

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

Your first consulting job

$$f: x \rightarrow y$$
$$P(Y|X)$$

- A billionaire from the suburbs of Seattle asks you a question:
 - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
 - You say: Please flip it a few times:



- You say: The probability is: .6
- **He says: Why???**
- You say: Because...

Thumbtack – Binomial Distribution

- $P(\text{Heads}) = \theta, P(\text{Tails}) = 1-\theta$

D : $\downarrow \downarrow \nwarrow \downarrow \nwarrow$
 $X_1 X_2 X_3 X_4 X_5$

$\theta \times \theta \times (1-\theta) \times \theta \times (1-\theta) = \theta^3 (1-\theta)^2$

identically distributed

- Flips are i.i.d.:

- ☐ Independent events
- ☐ Identically distributed according to Binomial distribution

- Sequence D of α_H Heads and α_T Tails



data likelihood $l(\theta) \rightarrow P(D | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
 - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Maximum Likelihood Estimate for Θ


$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$


- Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$$

■ Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0$$

$$\hat{\theta} = \arg \max_{\theta} \ln P(\mathcal{D} | \theta)$$

$$= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \stackrel{\partial = d}{=} \frac{d}{d\theta} \alpha_H \ln \theta + \alpha_T \ln(1 - \theta)$$

$$\frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}$$


$$\alpha_H \frac{\partial \ln \theta}{\partial \theta} + \alpha_T \frac{\partial \ln(1 - \theta)}{\partial \theta}$$

$$\alpha_H \frac{1}{\theta} + \alpha_T \frac{\frac{\partial(1 - \theta)}{\partial \theta} \cdot \frac{\partial \ln(1 - \theta)}{\partial(1 - \theta)}}{\partial \theta}$$

$$0 = \alpha_H \frac{1}{\theta} + \alpha_T (-1) \frac{1}{1 - \theta}$$

$$\theta = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

How many flips do I need?


$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Maximum
likelihood
Estimate.

Bayesian Learning

$$MLE = \underset{\theta}{\operatorname{argmax}} \underbrace{P(\mathcal{D} | \theta)}$$

- Use Bayes rule:

$$\underset{\theta}{\operatorname{argmax}} \underbrace{P(\theta | \mathcal{D})} = \frac{P(\mathcal{D} | \theta) P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

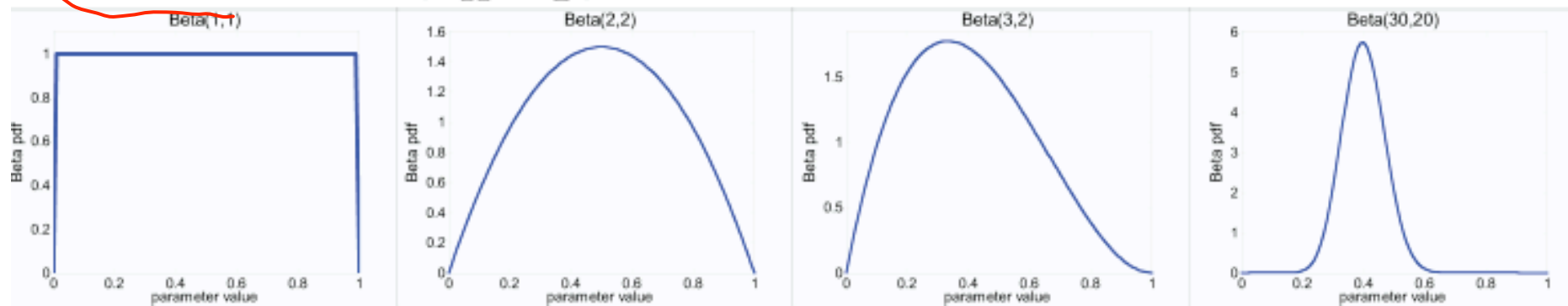
$$\underbrace{P(\theta | \mathcal{D})} \propto P(\mathcal{D} | \theta) P(\theta)$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Mean:

Mode:



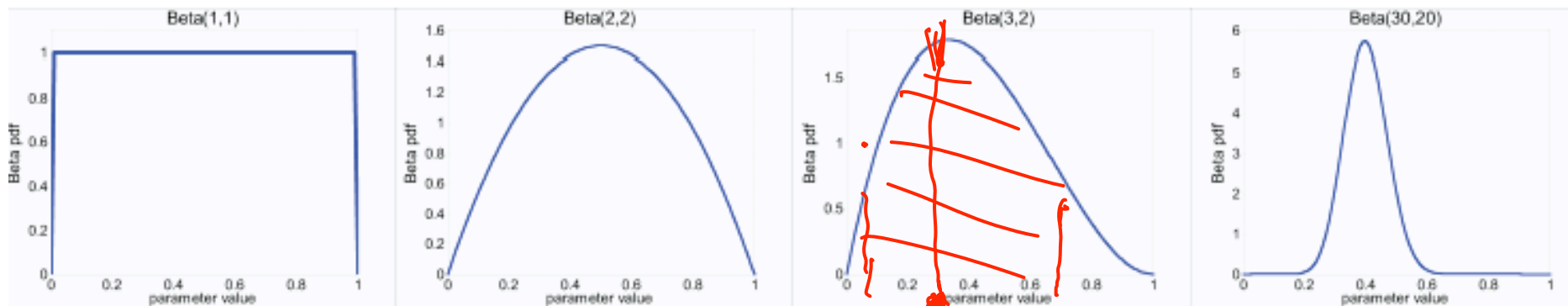
- Likelihood function: $P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

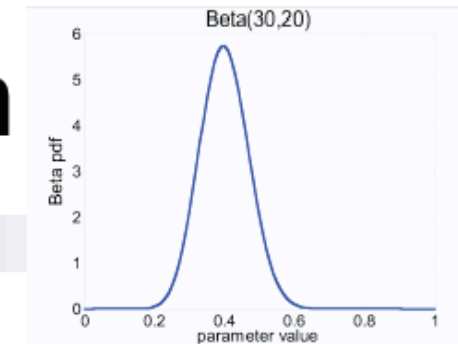
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



MAP est

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H}{\alpha_H + \alpha_T + \beta_H + \beta_T}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

Dirichlet distribution



- number of heads in N flips of a two-sided coin
 - follows a binomial distribution
 - Beta is a good prior (conjugate prior for binomial)
- what if it's not two-sided, but k-sided?
 - follows a multinomial distribution
 - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \dots, \theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_i - 1)}$$

Lejeune Dirichlet



Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	 Germany
Nationality	 German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data \mathcal{D}

$$\hat{\theta} = \arg \max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg \max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}\end{aligned}$$

You should know

- Probability basics
 - random variables, events, sample space, conditional probs, ...
 - independence of random variables
 - Bayes rule
 - Joint probability distributions
 - calculating probabilities from the joint distribution
- Point estimation
 - maximum likelihood estimates
 - maximum a posteriori estimates
 - distributions – binomial, Beta, Dirichlet, ...



Example: Bernoulli model

- Data:

- We observed N iid coin tossing: $D=\{1, 0, 1, \dots, 0\}$

- Representation:

Binary r.v:

$$x_n = \{0,1\}$$

- Model:

$$P(x) = \begin{cases} 1-\theta & \text{for } x=0 \\ \theta & \text{for } x=1 \end{cases} \Rightarrow P(x) = \theta^x (1-\theta)^{1-x}$$

- How to write the likelihood of a single observation x_i ?

$$P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}$$

- The likelihood of dataset $D=\{x_1, \dots, x_N\}$:

$$P(x_1, x_2, \dots, x_N | \theta) = \prod_{i=1}^N P(x_i | \theta) = \prod_{i=1}^N (\theta^{x_i} (1-\theta)^{1-x_i}) = \theta^{\sum_{i=1}^N x_i} (1-\theta)^{\sum_{i=1}^N 1-x_i} = \theta^{\text{\#head}} (1-\theta)^{\text{\#tails}}$$

