So...

… why do we keep having this debate:
rules/symbols vs. prototypes/connections?

So...

The real problem: a spurious contest between logic and probability.

- Neither logic nor probability on its own is sufficient to account for human cognition:
  - Generativity
  - Systematicity
  - Recursion and abstraction
  - Flexibility
  - Effective under great uncertainty (e.g., sparse data)
- What we really need is to understand how logic and probability can work together.

So…

The real problem: a spurious contest between logic and probability.

- A confusion between knowledge representations and inference processes:
  - Gradedness or fuzziness doesn't necessarily mean that the knowledge representations lack structure or rules -- merely that the inference processes incorporate uncertainty.

- Probabilistic inference over structured representations is what we need.

So...

… why do we keep having this debate: rules/symbols vs. prototypes/connections?

… why has it taken Cognitive Science much longer to get over it than AI?

Introduction to Bayesian inference

Representativeness in reasoning

Which sequence is more likely to be produced by flipping a fair coin?

HHTHT

HHHHH

A reasoning fallacy

Kahneman & Tversky: people judge the probability of an outcome based on the extent to which it is representative of the generating process.

Not wired for probability?

- Slovic, Fischhoff, and Lichtenstein (1976):
  - "It appears that people lack the correct programs for many important judgmental tasks.... it may be argued that we have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty."
- Gould (1992):
  - "Our minds are not built (for whatever reason) to work by the rules of probability."

Aristotle (4th century B.C.)

- In *On the heavens*, Aristotle asks whether the stars move independently or whether they are all fixed to some sphere.
- He observes that stars moving in large circles (near the celestial equator) take the same time to rotate as those near the polestar, which rotate in small circles.
- Infers a common cause: "If, on the other hand, the arrangement was a chance combination, the coincidence in every case of a greater circle with a swifter movement of the star contained in it is too much to believe. In one or two cases, it might not inconceivably fall out so, but to imagine it in every case alike is a mere fiction. Besides, chance has no place in that which is natural, and what happens everywhere and in every case is no matter of chance."

Image removed due to copyright considerations. Please see:

Halley. "Motuum Cometarum in Orbe Parabolico Elementa Astronomica."
In "Astronomiae Cometiae Synopsis." *Philisophical Transactions* (1705).

Transcript available at http://www.seds.org/~spider/spider/Comets/halley_p.html

Image removed due to copyright considerations. Please see:

Halley. "Motuum Cometarum in Orbe Parabolico Elementa Astronomica."
In "Astronomiae Cometiae Synopsis." *Philisophical Transactions* (1705).

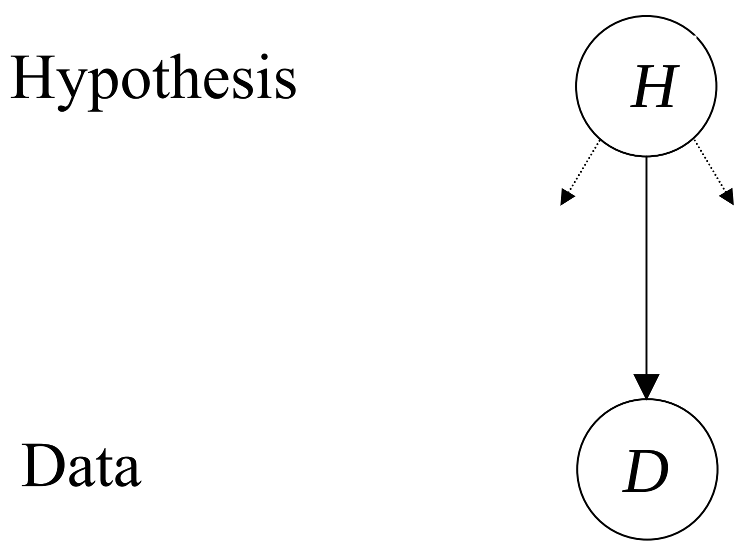
Transcript available at http://www.seds.org/~spider/spider/Comets/halley_p.html

A reasoning fallacy

Kahneman & Tversky: people judge the probability of an outcome based on the extent to which it is representative of the generating process.

But how does “representativeness” work?

Predictive versus inductive reasoning



Predictive versus inductive reasoning

<u>Prediction</u>

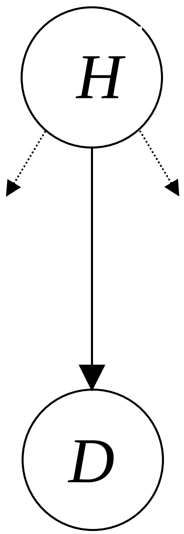
given

|

Likelihood: $P(D \mid H)$

↓

?



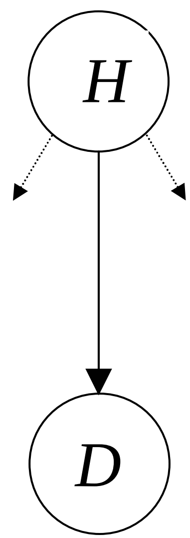
Predictive versus inductive reasoning

<u>Prediction</u>

given

Likelihood: $P(D \mid H)$

?



<u>Induction</u>

?

Representativeness

given

Bayes' rule

For data *D* and a hypothesis *H*, we have:

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)}$$

- "Posterior probability": $P(H \mid D)$
- "Prior probability": $P(H)$
- "Likelihood": $P(D \mid H)$

The origin of Bayes' rule

- A simple consequence of using probability to represent degrees of belief
- For any two random variables:

$$P(A \wedge B) = P(A)\, P(B \mid A)$$

$$P(A \wedge B) = P(B)\, P(A \mid B)$$

$$P(B)\, P(A \mid B) = P(A)\, P(B \mid A)$$

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}$$

Why represent degrees of belief with probabilities?

- Cox Axioms
  - necessary to cohere with common sense
- "Dutch Book" + Survival of the Fittest
  - if your beliefs do not accord with the laws of probability, then you can always be out-gambled by someone whose beliefs do so accord.
- Provides a theory of learning
  - a common currency for combining prior knowledge and the lessons of experience.

Cox Axioms (via Jaynes)

- Degrees of belief are represented by real numbers.
- Qualitative correspondence with common sense,

  e.g.: $Bel(\neg A) = f[Bel(A)]$

  $$Bel(A \wedge B) = g[Bel(A), Bel(B \mid A)]$$

- Consistency:
  - If a conclusion can be reasoned in more than one way, then every possible way must lead to the same result.
  - All available evidence should be taken into account when inferring a degree of belief.
  - Equivalent states of knowledge should be represented with equivalent degrees of belief.
- Accepting these axioms implies *Bel* can be represented as a probability measure.

Probability as propositional logic with uncertainty

- All of probability theory can be derived from these two laws (plus propositional logic):

$$P(A \mid I) + P(\neg A \mid I) = 1$$

$$P(A \wedge B \mid I) \equiv P(A, B \mid I) = P(A \mid B, I) \times P(B \mid I)$$

- That’s good: simple, elegant principles.
- That’s bad: how to work with structured representations? *More on that later….*

Bayesian inference

- Bayes' rule: $P(H \mid D) = \dfrac{P(H)P(D \mid H)}{P(D)}$

- What makes a good scientific argument? $P(H|D)$ is high if:
  - Hypothesis is plausible: $P(H)$ is high
  - Hypothesis strongly predicts the observed data:

    $P(D|H)$ is high
  - Data are surprising: $P(D)$ is low

A more useful form of Bayes

- Random variable *X* denotes a set of mutually exclusive exhaustive propositions (states of the world): $X = \{x_1, \ldots, x_n\}$

$$\sum_i P(X = x_i) = 1$$

- A useful rule: conditionalization

$$P(X = x_i) = \sum_j P(X = x_i \mid Y = y_j)\, P(Y = y_j)$$

A more useful form of Bayes

- Random variable *X* denotes a set of mutually exclusive exhaustive propositions (states of the world): $X = \{x_1, \ldots, x_n\}$

$$\sum_i P(X = x_i) = 1$$

- Bayes' rule for more than two hypotheses:

$$P(H = h \mid D = d) = \frac{P(H = h)P(D = d \mid H = h)}{P(D = d)}$$

A more useful form of Bayes

- Random variable *X* denotes a set of mutually exclusive exhaustive propositions (states of the world): $X = \{x_1, \ldots, x_n\}$

$$\sum_i P(X = x_i) = 1$$

- Bayes' rule for more than two hypotheses:

$$P(H = h \mid D = d) = \frac{P(H = h)P(D = d \mid H = h)}{\sum_i P(H = h_i)P(D = d \mid H = h_i)}$$

A more useful form of Bayes

- Random variable *X* denotes a set of mutually exclusive exhaustive propositions (states of the world): $X = \{x_1, \ldots, x_n\}$

$$\sum_i P(X = x_i) = 1$$

- Bayes' rule for more than two hypotheses:

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{\sum_i P(h_i)P(d \mid h_i)}$$

Sherlock Holmes

- “How often have I said to you that when you have eliminated the impossible whatever remains, however improbable, must be the truth?” (*The Sign of the Four*)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{\sum_i P(h_i)P(d \mid h_i)}$$

Sherlock Holmes

- "How often have I said to you that when you have eliminated the impossible whatever remains, however improbable, must be the truth?" (*The Sign of the Four*)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{P(h)P(d \mid h) + \sum_{h_i \neq h} P(h_i)P(d \mid h_i)}$$

Sherlock Holmes

- "How often have I said to you that when you have eliminated the impossible whatever remains, however improbable, must be the truth?" (*The Sign of the Four*)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{P(h)P(d \mid h) + \sum_{h_i \neq h} P(h_i)\boxed{P(d \mid h_i)}}$$

$= 0$

Sherlock Holmes

- "How often have I said to you that when you have eliminated the impossible whatever remains, however improbable, must be the truth?" (*The Sign of the Four*)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{P(h)P(d \mid h)} = 1$$

Sherlock Holmes

- "How often have I said to you that when you have eliminated the impossible whatever remains, however improbable, must be the truth?" (*The Sign of the Four*)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{\boxed{P(h)P(d \mid h)}} = 1$$

$$> 0$$

A reasoning fallacy

Kahneman & Tversky: people judge the probability of an outcome based on the extent to which it is representative of the generating process.

Hypotheses in coin flipping

Describe processes by which D could be generated

$$D = \mathsf{HHTHT}$$

- Fair coin, $P(\mathsf{H}) = 0.5$
- Coin with $P(\mathsf{H}) = \theta$
- Markov model
- Hidden Markov model
- ...

generative models

Representing generative models

- Graphical model notation
  - Pearl (1988), Jordan (1998)
- Variables are nodes, edges indicate dependency
- Directed edges show causal process of data generation

$$d_1 \quad d_2 \quad d_3 \quad d_4$$

Fair coin: $P(\mathsf{H}) = 0.5$

$$d_1 \rightarrow d_2 \rightarrow d_3 \rightarrow d_4$$

Markov model:

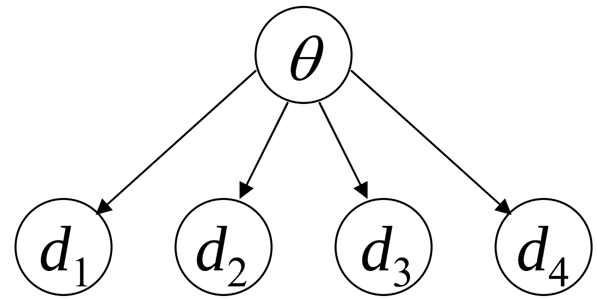
$$\begin{aligned} P(d_{i+1}|d_i) &= 0.7 \text{ if } d_{i+1} \neq d_i \\ &= 0.3 \text{ if } d_{i+1} = d_i \end{aligned}$$

HHTHT

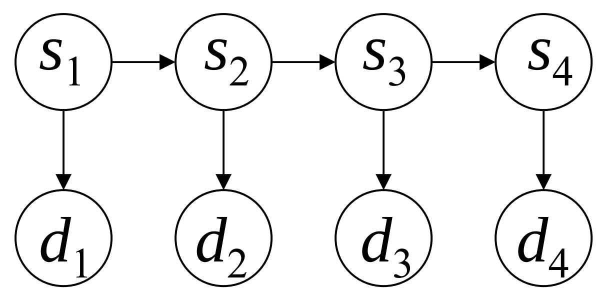
d_1 d_2 d_3 d_4 d_5

Models with latent structure

- Not all nodes in a graphical model need to be observed
- Some variables reflect *latent* structure, used in generating D but unobserved



$P(\mathsf{H}) = \theta$



Hidden Markov model:

$s_i \in$ {Fair coin, Trick coin}

HHTHT

$d_1 \; d_2 \; d_3 \; d_4 \; d_5$

Coin flipping

- Comparing two simple hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = 1.0$
- Comparing simple and complex hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = \theta$
- Comparing infinitely many hypotheses
  - $P(\mathsf{H}) = \theta$: Infer θ

Coin flipping

- Comparing two simple hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = 1.0$
- Comparing simple and complex hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = \theta$
- Comparing infinitely many hypotheses
  - $P(\mathsf{H}) = \theta$: Infer θ

Coin flipping

HHTHT

HHHHH

What process produced these sequences?

Comparing two simple hypotheses

- Contrast simple hypotheses:
  - H_1: "fair coin", $P(\mathsf{H}) = 0.5$
  - H_2:"always heads", $P(\mathsf{H}) = 1.0$
- Bayes' rule:

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)}$$

- With two hypotheses, use odds form

Bayes' rule in odds form

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	data
H_1, H_2:	models
$P(H_1\|D)$:	posterior probability H_1 generated the data
$P(D\|H_1)$:	likelihood of data under model H_1
$P(H_1)$:	prior probability H_1 generated the data

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHTHT		
H_1, H_2:	"fair coin", "always heads"		
$P(D|H_1) =$	$1/2^5$	$P(H_1) =$	?
$P(D|H_2) =$	0	$P(H_2) =$	1-?

$$P(H_1|D) \,/\, P(H_2|D) = \text{infinity}$$

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHTHT		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1) =$	$1/2^5$	$P(H_1) =$	999/1000
$P(D\|H_2) =$	0	$P(H_2) =$	1/1000

$$P(H_1|D) \;/\; P(H_2|D) = \text{infinity}$$

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHHHH		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1) =$	$1/2^5$	$P(H_1) =$	999/1000
$P(D\|H_2) =$	1	$P(H_2) =$	1/1000

$$P(H_1|D) \,/\, P(H_2|D) \approx 30$$

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHHHHHHHHH		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1) =$	$1/2^{10}$	$P(H_1) =$	999/1000
$P(D\|H_2) =$	1	$P(H_2) =$	1/1000

$$P(H_1|D) \,/\, P(H_2|D) \approx 1$$

The role of theories

The fact that HHTHT looks representative of a fair coin and HHHHH does not reflects our implicit theories of how the world works.

- Easy to imagine how a trick all-heads coin could work: high prior probability.
- Hard to imagine how a trick "HHTHT" coin could work: low prior probability.