Bayesianism

Christian Wüthrich

http://philosophy.ucsd.edu/faculty/wuthrich/

145 Philosophy of Science

Bayesian epistemology/Bayesian confirmation theory

- one of the most important developments in epistemology of C20
- offers a formal and mathematically rigorous framework of relating beliefs in hypotheses and evidence confirming or disconfirming it
- framework is probabilistic: assigns probabilities to beliefs
- General idea: a piece of evidence e confirms a hypothesis h in case it raises the probability of h, i.e., $P(h|e) > P(h)$
- probabilities should be 'updated' in a way predicted by Bayes's theorem, such that updated degree of belief in hypothesis is probability of hypothesis conditional on evidence
- includes a pragmatic 'self-defeat test' for epistemic rationality (next best thing to justification based on deductive logic)

⇒ laws of probability calculus as constraints on rational degrees of belief (or of confidence)

Andrey Nikolaevich Kolmogorov (1903-1987)

- Russian mathematician, Moscow State U
- contributions in probability thy, topology, intuitionistic logic, turbulence, classical mechanics, computational complexity
- main accomplishment: axiomatic foundation of probability theory
- “The theory of probability as mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra.”

Kolmogorov axioms

Given: class $\mathcal{S}$ of propositions $a, b, c, \ldots$

Introduce probability function on $\mathcal{S}$ as map from $\mathcal{S}$ to the closed interval $[0, 1]$ such that the following axioms hold:

Axiom (1: non-negativity)

$P(x) \geq 0$ for all x in $\mathcal{S}$; i.e. all probabilities are non-negative.

Axiom (2: unit measure)

$P(x) = 1$ if x in $\mathcal{S}$ is a tautology; i.e. if x is a proposition that is true in all possible cases, then it has a probability of 1.

Axiom (3: additivity)

For all x and y in $\mathcal{S}$, if x and y are mutually exclusive propositions, then $P(x \vee y) = P(x) + P(y)$.

Conditional probability

Definition (Conditional probability)

Given a probability function $P(x)$ as defined on the previous slide, the conditional probability $P(s|t)$ *of s given t is defined as*

$$P(s|t) := \frac{P(s\&t)}{P(t)}.$$

Bayesian updating of beliefs

Bayesians make (a more complicated version of) the following epistemological assumption

Principle (Conditionalization)

Starting from initial (prior) probabilities $P_i(h)$ of any hypothesis statement h, acquiring new evidence in the sense of becoming certain of the evidence statement e, rationality dictates that one updates one's initial probabilities to obtain one's final (posterior) probabilities by 'conditionalizing' on e

$$P_i(h) \rightarrow P_f(h) = P_i(h|e)$$

So we should find way to calculate $P_i(h|e)$ as a function of $P_i(h)$.

Theorem by Rev Thomas Bayes (1702-1761)

For all propositions h and e, we have

$$P(h|e) = \frac{P(h) \cdot P(e|h)}{P(e)} = \frac{P(h) \cdot P(e|h)}{P(e|h) \cdot P(h) + P(e|\neg h) \cdot P(\neg h)}$$

where

- $P(h)$: prior probability of h
- $P(h|e)$: posterior probability of h (in the light of e)
- $P(e|h)$: 'likelihood' of evidence e on hypothesis h

Bayesian updating

- first, determine the prior probability of h and the likelihood that e_1 will be observed given h
- determine the probability to observe e_1 *independently of h*
- if e_1 is observed, calculate the posterior probability $P(h|e_1)$ via Bayes's theorem
- consider this posterior probability as your new prior probability of h
- consider the probability of a new piece of evidence e_2 and its likelihood in the light of h
- if e_2 is observed, calculate the new posterior probability of h via Bayes's theorem
- ...

Example 1: from which bowl is the cookie?

Two bowls of cookies:

1. $Bowl_1$ has 10 chocolate chip and 30 plain cookies
2. $Bowl_2$ has 20 chocolate chip and 20 plain cookies

Question: If you pick a random cookie from a random bowl, and it is plain (e), how probable is it that it's from $Bowl_1$ (h)?

Priors: $P(h) = P(\neg h) = 0.5$

Likelihoods: $P(e|h) = 0.75$ and $P(e|\neg h) = 0.5$

Use Bayes's theorem:

$$\begin{aligned} P(h|e) &= \frac{P(h) \cdot P(e|h)}{P(e|h) \cdot P(h) + P(e|\neg h) \cdot P(\neg h)} \\ &= \frac{0.5 \cdot 0.75}{0.5 \cdot 0.75 + 0.5 \cdot 0.5} = 0.6. \end{aligned}$$

Example 2: is she/he going to the party?

from Peter Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science*, p. 204.

- h: hypothesis that she/he is at party
- e: evidence that her/his car is parked outside
- $P(h)$: initial probability that she/he is at party (before seeing the car); let's say this is 0.5.
- $P(e|h)$: likelihood that her/his car is parked outside if she/he is at the party; suppose this is 0.8
- $P(e|\neg h)$: likelihood that her/his car is parked outside if she/he is not at the party; suppose this is only 0.1
- $P(h|e)$: prob that she/he is at the party given that her/his car is parked outside; can be calculated using Bayes's theorem:

$$P(h|e) = \frac{0.5 \cdot 0.8}{0.5 \cdot 0.8 + 0.5 \cdot 0.1} = 0.89.$$

$\Rightarrow$ seeing the car raises the prob of h from 0.5 to 0.89.

Example 3: in the courtroom

Juror must assess how evidence bears on guilt of defendant:

- g: hypothesis that defendant is guilty
- e: evidence that defendant's DNA matches DNA found at crime scene
- $P(e|g)$: likelihood to see evidence of matching DNA if defendant is guilty; in capital offenses, typically very high; here assumed to be 1
- $P(e|\neg g)$: likelihood to see evidence of matching DNA if defendant is not guilty; very low, assume 1 in a million, or 10^{-6}
- $P(g)$: initial probability that defendant is guilty (prior); hugely depends on other evidence, circumstances etc. Two cases: either (A) strong prior suspicion ($P(g) = 0.3$), or (B) very low suspicion ($P(g) = 10^{-6}$)
- $P(g|e)$: prob that defendant is guilty if matching DNA is found; this is what we want to know!

Case (A): $P(g) = 0.3$

$$\begin{aligned} P(g|e) &= \frac{P(g) \cdot P(e|g)}{P(e|g) \cdot P(g) + P(e|\neg g) \cdot P(\neg g)} \\ &= \frac{0.3 \cdot 1.0}{0.3 \cdot 1.0 + 0.7 \cdot 10^{-6}} \\ &= 0.99999766667 \end{aligned}$$

Case (B): $P(g) = 10^{-6}$

$$\begin{aligned} P(g|e) &= \frac{P(g) \cdot P(e|g)}{P(e|g) \cdot P(g) + P(e|\neg g) \cdot P(\neg g)} \\ &= \frac{10^{-6} \cdot 1.0}{10^{-6} \cdot 1.0 + (1 - 10^{-6}) \cdot 10^{-6}} \\ &\approx 0.5 \end{aligned}$$

Example 4: the *Scorpion* search

- May 1968: US nuclear submarine *Scorpion* fails to arrive at home port of Norfolk, VA
- US Navy: convinced that vessel had been lost off Eastern seabord, but extensive search fails to discover wreck
- US Navy deep water expert John Craven believed that it was southwest of Portuguese archipelago of the *Azores* based on controversial triangulation of hydrophones
- allocation of limited resources (one ship) $\Rightarrow$ optimize them
- Craven worked with mathematicians to optimize the search, using Bayesian search theory
- October 1968: wreck is found 400 miles southwest of the Azores

Bayesian search theory

1. Sea is divided into grid squares
2. experienced submarine commanders are interviewed, etc, to formulate a number of hypotheses about what happened to vessel
3. construct prob distribution over squares corresponding to each hypothesis
4. construct prob distribution for actually finding object in square X if it really is in X (function of water depth)
5. combine all these prob distributions (from 3 and 4) to produce overall probability grid; this gives prob of finding object in a square if this square is searched (for all squares)
6. construct a search path starting from square of highest prob that then searches high prob areas, then intermediate prob areas, then low prob areas
7. revise overall prob distribution continuously as you search, i.e. if you have unsuccessfully searched square, then prob that object is there is greatly reduced (though usually not zero), and prob of finding it elsewhere must be increased; this revision follows Bayes's theorem

Example 5: V1@gra and Bayesian filtering

Introduction to Bayesian filtering

- task: figure out how likely an email message is spam based on the words that appear in it
- basic idea: Bayesian spam filter
  - lists words in incoming emails,
  - assigns to each word probability that it appears in spam mail (misspelled words score very high), and
  - uses these probs as input into Bayes's formula to determine whether or not email is spam
- first need to train the spam filter by showing it spam and non-spam mails (more and more automated)
- spam filter stores all words in trained messages (incl host name, IP address, HTML tag, etc) in databases

- ⇒ filter calculates prob how likely it is that a word is in spam based on its frequency in databases (the 'spamicity' of each word)
- spamicity of 0.5 is neutral, higher (lower) means that it often occurs in (non-)spam messages
- filter then uses Bayes's formula to calculate the overall spamicity of a message based on the spamicity of all the words that occur in it
- ⇒ message put in spam filter if spamicity is above 0.5
- Generally: Bayesian spam filter are highly effective because (1) they adapt to individual circumstances (databases are built for each user), and (2) they learn over time and update the databases

Example 6: Predicting the US presidential election

How many electoral votes will Obama get? (in 2008)

- The link above leads to a blog where a simple Bayesian method is used to predict the outcome of the last US presidential election (on 31 October 2008).
- The blogger predicted, based on polling data from cnn.com and 5000 simulated elections, that the probability that McCain would win enough electoral votes to win him the White House was 0.0.
- Nate Silver made similar predictions for 2008. Here is his prediction for the 2012 presidential election based on Bayesian statistics the night before the election: http://fivethirtyeight.blogs.nytimes.com/fivethirtyeights-2012-forecast/
- And by Drew Linzer at Votamatic at http://votamatic.org/final-result-obama-332-romney-206/ (for more on Linzer's method: http://votamatic.org/how-it-works/)

What is probability?

(objective) chance	vs.	(subjective) degree of belief
$\downarrow$		$\downarrow$
How to measure?		How to measure?
$\downarrow$		$\downarrow$
frequency of occurrence		gambling behaviour (via wagers)

Axioms of probability must apply to both!

Subjectivist Bayesianism

- ‘probability as degree of personal belief’ (generated by free choice, socialization, evolution, etc)
- The probability of an event is just the certainty with which a Bayesian agent expects the event to occur.
- Main idea: ‘rational belief’ should be understood as a generalization of betting behaviour: given an amount of information/data and asked for an evaluation, what odds would one bet for the truth of one’s evaluation?
- to bet on h at odds of $X : 1$ is to be willing to risk losing \$$X$ if h is false, in return for a gain of \$1 if h is true
- if your subjectively fair odds for a bet on h are $X : 1$, then your degree of belief in h is $X/(X+1)$
- It is of course possible that the subjective degree of belief violates the axioms, but...

'Dutch book' theorem

If somebody's subjective degree of belief violates a Kolmogorov axioms, then this person should accept a combination of bets which amounts to a so-called Dutch book, i.e.

the combination of bets that should they be accepted guarantee the person a loss!

Simple example: your degree of belief that next coin toss will come out 'heads' is 0.55, and your degree of belief that it'll be 'tails' is 0.5 $\Rightarrow$ bookie can write out set of wagers which guarantee that you'll lose 5c on each dollar you bet

Decision table on combination of wages:

	$h(\neg t)$	$t(\neg h)$
bet \$0.55	\$1	\$0
bet \$0.50	\$0	\$1
pay \$1.05 in total	win \$1	win \$1

Bayesian solution to grue paradox

- Suppose you're presented two inductive arguments from the same set of observations of green emeralds, one arguing that all emeralds are green, the other that they are grue.
- Why is one induction better than the other?
- Standard Bayesian answer: both are OK, but most people would assign higher prior prob to 'green' hypothesis than to 'grue' hypothesis
- Reaction: true, it gives a difference, but does it explain why the 'grue' hypothesis gets lower prior prob?
- Bayesianism offers no criticism of subjective decision to assign high prior prob to 'grue' hypothesis, as long as probs are internally coherent and updated properly

(1) Problem of the priors

- initial set of prior probabilities can be chosen freely (except for 0 and 1)
- but how could a strange assignment of priors be criticized, so long as it follows the axioms?
- Bayesian answer: doesn't matter because initial set of priors are washed out asymptotically (convergence, stable estimation theorem)
- problem: conversely, we also have for any amount of evidence, and any measure of agreement, there is some set of priors s.t. this evidence will not get the two people to agree by the end (Kyburg)
- there must be agreement concerning the likelihoods $P(e_i|h)$, relevance of particular pieces of evidence
- assumptions of theorems do not even remotely apply in realistic scientific contexts

(2) Explanation of methodological truisms (Glymour)

- confirmation theory ought to explain general methodological truisms as well as particular judgments that have occurred in hist of science
- Bayesianism cannot account for
  - what makes a hypothesis ad hoc
  - what makes one body of evidence more various than another body of evidence
  - why we should prefer a variety of evidence
  - why we should prefer simpler theories (e.g. curve-fitting problem!)
- generally: Bayesianism cannot explain why we make the projections we make and why different evidence might be differently relevant (and why we might disagree in attributing this relevance)

(3) Problem of old evidence (Glymour)

- problem of old evidence: old evidence can in fact confirm new thy, but according to Bayesian kinematics it cannot
- suppose e is known before theory T is introduced at time t
- because e is known at t, $P_t(e) = 1$

$\Rightarrow$ likelihood of e given T is also 1: $P_t(e|T) = 1$

$$P_t(T|e) = \frac{P_t(T) \cdot P_t(e|T)}{P_t(e)} = P_t(T)$$

$\Rightarrow$ posterior prob of T is same as its prior prob!

By way of conclusion

Hájek and Hartmann, "Bayesian epistemology", *A Companion to Epistemology*, Oxford: Blackwell, 2009.

Alan Hájek and Stephan Hartmann contrast two views on Bayesian epistemology:

> *According to one view, there cannot [be a Bayesian epistemology]: Bayesianism fails to do justice to essential aspects of knowledge and belief, and as such it cannot provide a genuine epistemology at all. According to another view, Bayesianism should supersede traditional epistemology: where the latter has been mired in endless debates over skepticism and Gettierology, Bayesiansim offers the epistemologist a research program. We will advocate a more moderate view: Bayesianism can illuminate various long-standing problems of epistemology, while not addressing all of them; and while Bayesianism opens up fascinating new areas of research, it by no means closes down the staple preoccupations of traditional epistemology. (93)*