STSCI 4780:
Key theorems

Tom Loredo, CCAPS & DSS, Cornell University

2018-02-01

Recap: Probability theory as logic

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

$$
\begin{aligned}
P &= 0 \;\rightarrow\; \text{Argument is } \textit{invalid}\text{; premises imply } \overline{H} \\
&= 1 \;\rightarrow\; \text{Argument is } \textit{valid} \\
&\in (0,1) \;\rightarrow\; \text{Degree of deducibility}
\end{aligned}
$$

Mathematical model for induction

'AND' (product rule): $$\begin{aligned} P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P})\, P(B|A \wedge \mathcal{P}) \\ &= P(B|\mathcal{P})\, P(A|B \wedge \mathcal{P}) \end{aligned}$$

'OR' (sum rule): $$\begin{aligned} P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\ &\quad - P(A \wedge B|\mathcal{P}) \end{aligned}$$

'NOT': $$P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation*.

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful*, none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities*, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Harold Jeffreys (1931)

If we like there is no harm in saying that a probability expresses a degree of reasonable belief. . . . 'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. *Essentially the notion can only be described by reference to instances where it is used.* It is intended to express *a kind of relation between data and consequence* that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

More On Interpretation

Physics uses words drawn from ordinary language—mass, weight, momentum, force, temperature, heat, etc.—but their technical meaning is more abstract than their colloquial meaning. We can map between the colloquial and abstract meanings associated with specific values by using specific instances as "calibrators."

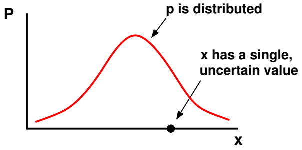
A Thermal Analogy

Intuitive notion	*Quantification*	*Calibration*
Hot, cold	Temperature, T	Cold as ice = 273K Boiling hot = 373K
uncertainty	Probability, P	Certainty = 0, 1 $p = 1/36$: plausible as "snake's eyes" $p = 1/1024$: plausible as 10 heads

Interpreting distributions

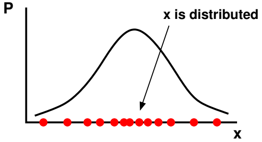
Bayesian

Probability *quantifies uncertainty* in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in the single case before us:



Frequentist

Probability *quantifies variability* across an ensemble, in terms of a rate/proportion/frequency. $p(x)$ describes how the *values of x* would be distributed among infinitely many trials:



Arguments Relating Hypotheses, Data, and Models

We seek to appraise scientific hypotheses in light of observed data and modeling assumptions.

Consider the data and modeling assumptions to be the premises of an argument with each of various hypotheses, H_i, as conclusions: $H_i|D_{\rm obs}, I$. ($I =$ "background information," everything deemed relevant besides the observed data)

$P(H_i|D_{\rm obs}, I)$ measures the degree to which $(D_{\rm obs}, I)$ allow one to deduce H_i. It provides an ordering among arguments for various H_i that share common premises.

Probability theory tells us how to analyze and appraise the argument, i.e., how to calculate $P(H_i|D_{\rm obs}, I)$ from simpler, hopefully more accessible probabilities.

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises

'AND' (product rule):
$$P(H_i, D_{\rm obs}|\mathcal{C}) = P(H_i|\mathcal{C})\, P(D_{\rm obs}|H_i, \mathcal{C}) = P(D_{\rm obs}|\mathcal{C})\, P(H_i|D_{\rm obs}, \mathcal{C})$$

'OR' (sum rule):
$$P(H_1 \vee H_2|\mathcal{C}) = P(H_1|\mathcal{C}) + P(H_2|\mathcal{C}) - P(H_1, H_2|\mathcal{C})$$

'NOT':
$$P(\overline{H_i}|\mathcal{C}) = 1 - P(H_i|\mathcal{C})$$

Three Important Theorems

Bayes's Theorem (BT)

Consider $P(H_i, D_{\text{obs}}|\mathcal{C})$ using the product rule:

$$\begin{aligned} P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C})\,P(D_{\text{obs}}|H_i,\mathcal{C}) \\ &= P(D_{\text{obs}}|\mathcal{C})\,P(H_i|D_{\text{obs}},\mathcal{C}) \end{aligned}$$

Solve for the *posterior probability* (expands the premises!):

$$P(H_i|D_{\text{obs}},\mathcal{C}) = \frac{P(H_i, D_{\text{obs}}|H_i,\mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})} = P(H_i|\mathcal{C})\,\frac{P(D_{\text{obs}}|H_i,\mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

posterior $\propto$ *prior* $\times$ *likelihood*

norm. const. $P(D_{\text{obs}}|\mathcal{C})$ = *prior predictive*

Aside: likelihood vs. probability

Data influence inferences via the ability of rival hypotheses to predict the actually observed data, quantified by $P(D_{\rm obs}|H_i, \mathcal{C})$

Consider

$$p(D|H_i, \mathcal{C})$$

This is a probability for choices of D, but not for choices of H_i (wrong side of the solidus!), although it does depend on H_i

For a particular choice of H_i, the resulting function of D specifies the *sampling distribution* or (more descriptively!) the *conditional predictive distribution* for data

The H_i dependence when we fix attention on the *observed* data is the *likelihood function* (for H_i, *not* for data):

$$\mathcal{L}(H_i) \equiv p(D_{\rm obs}|H_i, \mathcal{C})$$

The likelihood for H_i is not a probability for H_i, but the posterior probability for H_i is proportional to it

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ($\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A, \mathcal{C}) P(A|\mathcal{C}) = P(A|\mathcal{C}) \\ &= \sum_i P(B_i|\mathcal{C}) P(A|B_i, \mathcal{C})\end{aligned}$$

If we do not see how to get $P(A|\mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{C}) P(A|B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Example: Take $\mathcal{P} = \mathcal{C}$, $A = D_{\rm obs}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\rm obs}|\mathcal{C}) &= \sum_i P(D_{\rm obs}, H_i|\mathcal{C}) \\ &= \sum_i P(H_i|\mathcal{C})P(D_{\rm obs}|H_i, \mathcal{C}) \end{aligned}$$

prior predictive for $D_{\rm obs}$ = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Normalization

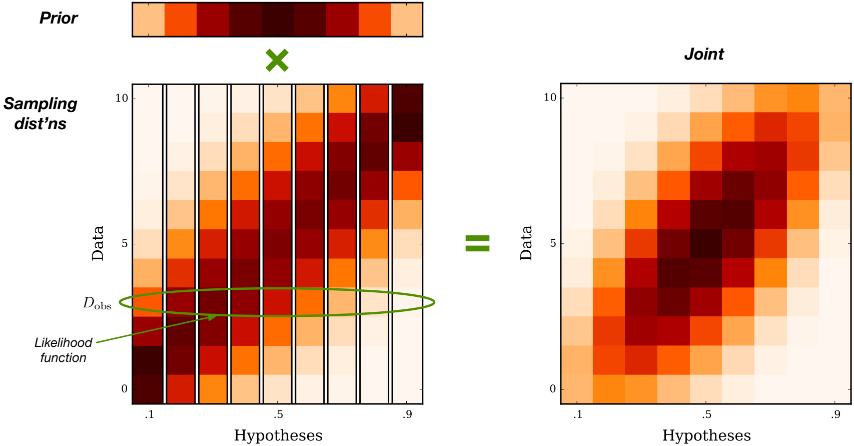
For *exclusive, exhaustive* H_i,

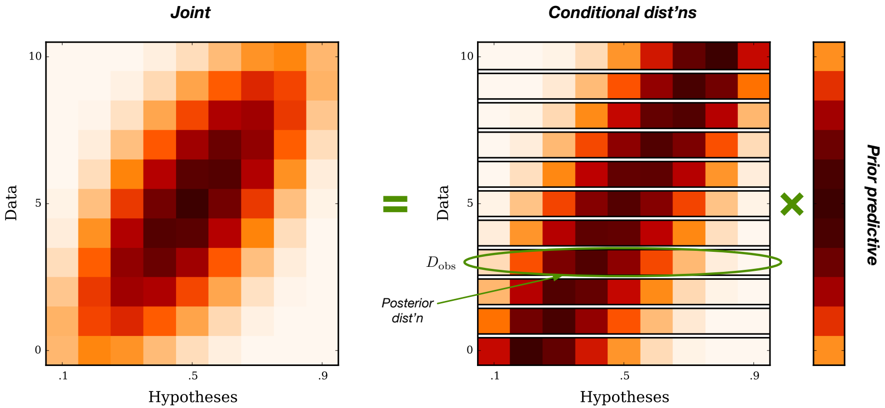
$$\sum_i P(H_i|\cdots) = 1$$

Inference as manipulation of the joint distribution

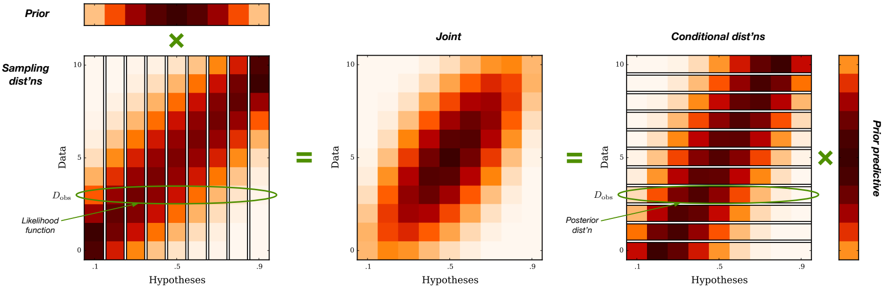
Bayes's theorem in terms of the *joint distribution*:

$$P(H_i|\mathcal{C}) \times P(D_{\text{obs}}|H_i, \mathcal{C}) = P(H_i, D_{\text{obs}}|\mathcal{C}) = P(H_i|D_{\text{obs}}, \mathcal{C}) \times P(D_{\text{obs}}|\mathcal{C})$$




Joint
Conditional dist'ns
Data
10
5
0
.1
.5
.9
Hypotheses
=
D_{obs}
Posterior dist'n
×
Prior predictive

$$P(H_i|\mathcal{C}) \times P(D_{\text{obs}}|H_i,\mathcal{C}) = P(H_i, D_{\text{obs}}|\mathcal{C}) = P(H_i|D_{\text{obs}},\mathcal{C}) \times P(D_{\text{obs}}|\mathcal{C})$$



Well-Posed Problems

The rules express desired probabilities in terms of other probabilities

To get a numerical value *out*, at some point we have to put numerical values *in*

Direct probabilities are probabilities with numerical values determined directly by premises (via modeling assumptions, symmetry arguments, previous calculations, desperate presumption . . .)

An inference problem is *well posed* only if all the needed probabilities are assignable based on the context. We may need to add new assumptions as we see what needs to be assigned. We may not be entirely comfortable with what we need to assume! (Remember Euclid's fifth postulate!)

Should explore how results depend on uncomfortable assumptions ("robustness")

Simplest case: Binary classification

Context:

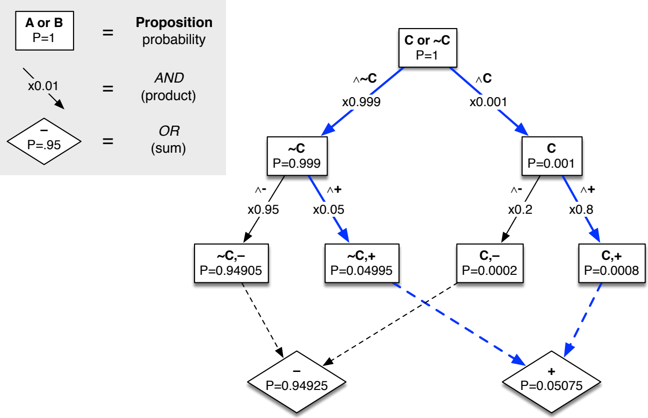
- 2 hypotheses: $\{C, \overline{C}\}$
- 2 possible data values: $\{-, +\}$
- Info about test accuracy

Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(\overline{C})$?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
    - Test is 80% accurate if you have the condition: $P(+|C, \mathcal{C}) = 0.8 \quad$ ("sensitivity")
    - Test is 95% accurate if you are healthy: $P(-|\overline{C}, \mathcal{C}) = 0.95 \quad$ ("specificity," $1 - p(\text{false } +)$)

Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women

Case diagram—probabilities



$$P(H_1 \vee H_2|\mathcal{C})$$

$$P(H_i|\mathcal{C})$$

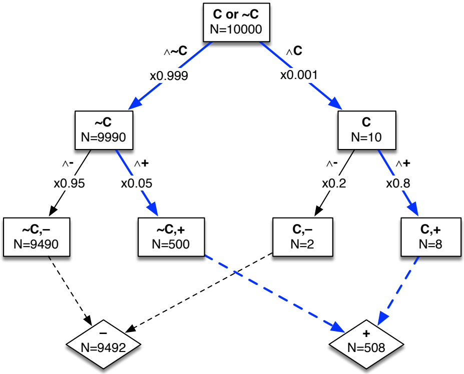
$$P(H_i, D_j|\mathcal{C}) = P(H_i|\mathcal{C})\, P(D_j|H_i, \mathcal{C})$$

$$P(D_j|\mathcal{C}) = \sum_i P(H_i, D_j|\mathcal{C})$$

$$P(C|+, \mathcal{C}) = \frac{0.0008}{0.05075} \approx 0.016$$

Case diagram—counts

Create a large ensemble of imaginary cases so ratios of counts approximate the probabilities.



$$P(C|+,\mathcal{C}) = \frac{8}{508} \approx 0.016$$

Of the 508 cases with positive test results, only 8 have the condition. The prevalence is so low that when there is a positive result, it's more likely to have been a mistake than accurate, even for a sensitive test.