Pearl's & Jeffrey's update rules in probabilistic learning

Radboud University Nijmegen
CCHM, Oxford, Jan. 12, 2024

Bart Jacobs
bart@cs.ru.nl

Outline

Pearl's & Jeffrey's update

Where we are, so far

Challenges in probabilistic logic (from Pearl'89)

> *To those trained in traditional logics, symbolic reasoning is the standard, and nonmonotonicity a novelty. To students of probability, on the other hand, it is symbolic reasoning that is novel, not nonmonotonicity. Dealing with new facts that cause probabilities to change abruptly from very high values to very low values is a commonplace phenomenon in almost every probabilistic exercise and, naturally, has attracted special attention among probabilists. The new challenge for probabilists is to find ways of abstracting out the numerical character of high and low probabilities, and cast them in linguistic terms that reflect the natural process of accepting and retracting beliefs.*

Embarrassingly, there is still **no** probabilistic logic for symbolic reasoning.

Probabilistic reasoning and updating (belief revision)

Example

I may think that scientists are civilised people. But then I attend a conference dinner that ends in a fist fight.

I will update my judgement.

- This is difficult in traditional, monotonic logic, where adding more information can not make true statements false.
- We need to switch from truth/falsity of statement, to likelihood
  - not two-element set $\{0, 1\}$ but interval $[0, 1]$
  - not sharp but fuzzy (soft) statements

  The likelihood that scientists are civilised is decreased, by the events at the conference dinner, through updating (belief revision).

Naive picture of learning

"Nürnberger Trichter"
(Nurnberg Funnel)

Alternative: predictive coding theory (Karl Friston et al)

- The human mind is constantly active in making predictions
- These predictions are compared with what actually happens
- Mismatches (prediction errors) lead to updates in the brain

> "The human brain is a Bayesian prediction & correction engine"

Possibly it is better to call the mind a Jeffreyan engine ...

My own (logical) interests/work

- There are two update rules, by Judea Pearl (1936) and by Richard Jeffrey (1926-2002), which are not well-distinguished in the literature
  - They both have clear formulations using channels — see later
  - What are the differences? When to use which rule? Unclear!
- The topic is mathematically non-trivial
  - esp. in Jeffrey's case, as we shall see
- Intriguing question: does the human mind use Pearl's or Jeffrey's rule — within predictive coding theory
  - cognitive science may provide an answer

- BJ, *The Mathematics of Changing one's Mind, via Jeffrey's or via Pearl's update rule*, Journ. of AI Research, 2019
- BJ, *Learning from What's Right and Learning from What's Wrong*, MFPS'21
- BJ & Dario Stein, *Pearl's and Jeffrey's Update as Modes of Learning in Probabilistic Programming*, MFPS'23

Example I, medical test, part I

- Consider a disease with *a priori* probability (or 'prevalence') of 10%
- There is a test for the disease with:
  - ('sensitivity') If someone has the disease, then the test is positive with probability of 90%
  - ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.
- Computing the predicted positive test probability yields: 13.5%
- The test is performed, under unfavourable circumstances like bad light, and we are only 80% sure that the test is positive. What is the disease likelihood?
- Updating with { **Pearl**'s rule gives: 26% disease likelihood; **Jeffrey**'s rule gives: 54%
- Jeffrey is more than twice as high as Pearl. Which should a doctor use?

Example II: multiple test results

In the same test set-up as before, you test three times and get:

two positive tests and one negative test

What is the posterior disease probability?

Updating with { **Pearl**'s rule gives: 79% disease likelihood; **Jeffrey**'s rule gives: 49%

Some remarks

- Computationally, Pearl's approach does not scale to many, many tests — unless there is a conjugate prior situation
- A possible interpretation for the difference:
  - Pearl is about tests for one individual
  - Jeffrey is about tests for a population, with different indivudals

Pearl & Jeffrey updating as optimisations

(What is formulated informally at this stage, will be made mathematcially precise later)

(1) **Pearl's rule:**
- uses evidence (predicate) to update a *prior* to a *posterior*
- such that the validity (expected value) of the evidence increases
- formally: the validity of the evidence in the prediction based on the posterior is higher than in the predication based on the prior

(2) **Jeffrey's rule:**
- uses an observed distribution/state to update from *prior* to *posterior*
- such that the mismatch with the observation decreases
- formally: the KL-divergence between the observation and the prediction based on the posterior is lower than on the prior

Thus, Jeffrey's rule reduces prediction errors, as in predictive coding

Where we are, so far

Comparison table about updating (with informal descriptions)

	Pearl's rule	Jeffrey's rule
effect	increase of what's right	decrease of what's wrong
you learn nothing from	uniformity (no differences)	what you already know (predict)
successive updates commute?	yes	no

Big question

- Does the human mind use Pearl's or Jeffrey's rule?
- My bet is on Jeffrey ...
- Since the human mind is very sensitive to the order of updating (priming)

Where we are, so far

Distributions (finite, discrete)

A distribution (or state) over a set X is a formal finite convex sum:

$$\sum_i r_i |\, x_i \,\rangle \;\in\; \mathcal{D}(X) \qquad \text{where} \quad \begin{cases} r_i \in [0,1], \text{ with } \sum_i r_i = 1 \\ x_i \in X \end{cases}$$

- Distributions can also be described as functions $\sigma \colon X \to [0,1]$ with finite support and $\sum_x \sigma(x) = 1$
- This $\mathcal{D}$ is the distribution monad on Sets
- A Kleisli map $X \to \mathcal{D}(Y)$ is also called a channel, and written as $X \nrightarrow Y$, with special arrow. Channels capture conditional probabilities $p(Y|X)$ in a graphical calculus
- For $\sigma \in \mathcal{D}(X)$ and $c \colon X \nrightarrow Y$ we have Kleisli extension / bind / state transformation / prediction: $c \gg\!= \sigma \;\in\; \mathcal{D}(Y)$. Explicitly, if $\sigma = \sum_i r_i |\, x_i \,\rangle$, prediction along channel c is:

$$c \gg\!= \sigma \;:=\; \sum_i r_i \cdot c(x_i) \;=\; \sum_{y \in Y} \Big(\sum_i r_i \cdot c(x_i)(y) \Big) \,|\, y \,\rangle.$$

The disease-test example: state & channel

- Use sets $D = \{d, d^\perp\}$ for disease (or not) and $T = \{p, n\}$ for positive and negative test outcomes
- The prevalence state / distribution is:

$$prior = \tfrac{1}{10}|\,d\,\rangle + \tfrac{9}{10}|\,d^\perp\,\rangle.$$

- Testing is done via the channel $test\colon D \to \mathcal{D}(T)$ with:

$$test(d) = \tfrac{9}{10}|\,p\,\rangle + \tfrac{1}{10}|\,n\,\rangle \qquad \text{and} \qquad test(d^\perp) = \tfrac{1}{20}|\,p\,\rangle + \tfrac{19}{20}|\,n\,\rangle.$$

(Recall: sensitivity is $90\% = \frac{9}{10}$, specificity is $95\% = \frac{19}{20}$)

- The predicted test distribution is:

$$test \gg\!= prior = \tfrac{27}{200}|\,p\,\rangle + \tfrac{173}{200}|\,n\,\rangle = 0.135|\,p\,\rangle + 0.865|\,n\,\rangle.$$

This gives the 13.5% likelihood of positive tests.

Divergence between states

For $\omega, \rho \in \mathcal{D}(X)$ the Kullback-Leibler divergence, or *KL-divergence*, or simply *divergence*, of ω from ρ is:

$$D_{KL}(\omega, \rho) := \sum_{x \in X} \omega(x) \cdot \log\left(\frac{\omega(x)}{\rho(x)}\right).$$

It is one standard way to compare states.

Lemma (Basic divergence properties)

(1) $D_{KL}(\omega, \rho) \geq 0$, *with* $D_{KL}(\omega, \rho) = 0$ *iff* $\omega = \rho$
(2) *But:* $D_{KL}(\omega, \rho) \neq D_{KL}(\rho, \omega)$, *in general*
(3) *Also (but not used):* $D_{KL}\big(c \gg\!= \omega,\ c \gg\!= \rho\big) \leq D_{KL}(\omega, \rho)$
(4) *And:* $D_{KL}\big(\omega \otimes \omega',\ \rho \otimes \rho'\big) = D_{KL}(\omega, \rho) + D_{KL}(\omega', \rho')$

Predicates and transformations

A predicate on a set X is a function $p\colon X \to [0, 1]$.

- Each subset/event $E \subseteq X$ forms a 'sharp' predicate, via the indicator function $1_E\colon X \to [0, 1]$
- For each $x \in X$ write $1_x = 1_{\{x\}}$ for the point predicate, sending $x' \neq x$ to 0 and x to 1.

Given a channel $c\colon X \nrightarrow Y$ and a predicate q on Y, one defines predicate transformation $c \mathrel{=\!\!\ll} q$, as predicate on X.
Explicitly, on $x \in X$,

$$\Big(c \mathrel{=\!\!\ll} q\Big)(x) := \sum_{y \in Y} c(x)(y) \cdot q(y).$$

Note: state tranformation $\gg\!=$ goes in forward direction, along the channel, and predicate transformation $=\!\!\ll$ goes backward.

Validity and conditioning

(1) For a state ω on a set X, and a predicate p on X define validity as:

$$\omega \models p \quad := \quad \sum_{x \in X} \omega(x) \cdot p(x) \quad \in \quad [0, 1]$$

It describes the expected value of p in ω.

(2) If $\omega \models p$ is non-zero, we define the conditional distribution $\omega|_p$ as:

$$\omega|_p(x) := \frac{\omega(x) \cdot p(x)}{\omega \models p} \qquad \text{that is} \qquad \omega|_p = \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} |\,x\,\rangle.$$

It's the normalised product of ω and p.

Two basic results about validity $\models$

Theorem (Validity and transformation)

For channel $c\colon X \nrightarrow Y$, state σ on X, predicate q on Y,

$$c \gg= \sigma \models q \quad = \quad \sigma \models c \ll= q$$

Theorem (Validity increase)

For a state ω and predicate p (on the same set, with non-zero validity),

$$\omega|_p \models p \quad \geq \quad \omega \models p$$

Informally, absorbing evidence p into state ω, makes p more true.

The "dagger" of a channel: Bayesian inversion

Assume a channel $c\colon X \nrightarrow Y$ and a state $\sigma \in \mathcal{D}(X)$.

- For an element $y \in Y$ we can form:
  (1) the point predicate 1_y on Y
  (2) its transformation $c \ll= 1_y$ along c, as predicate on X
  (3) the updated state $\sigma|_{c \ll= 1_y} \in \mathcal{D}(X)$.
- This yields an inverted channel, the "dagger"

$$Y \xrightarrow{c^\dagger_\sigma} X \qquad \text{with} \qquad c^\dagger_\sigma(y) := \sigma|_{c \ll= 1_y}$$

- This forms a dagger functor on a symmetric monoidal category.
  - see e.g. Clerc, Dahlqvist, Danos, Garnier in FoSSaCS 2017
  - with disintegration: Cho-Jacobs in MSCS'19; Fritz in AIM'20.

Pearl and Jeffrey, formulated via channels (JAIR'19)

Set-up:

- a channel $c\colon X \nrightarrow Y$ with a (prior) state $\sigma \in \mathcal{D}(X)$ on the domain
- evidence on Y, that we wish to use to update σ

- **Pearl's update rule**
  (1) Evidence is a predicate q on Y
  (2) Updated state:
  $$\sigma_P := \sigma|_{c \ll= q}$$
- **Jeffrey's update rule**
  (1) Evidence is state τ on Y
  (2) Updated state:
  $$\sigma_J := c^\dagger_\sigma \gg= \tau = \sum_{y \in Y} \tau(y) \cdot \left(\sigma|_{c \ll= 1_y}\right)$$

Main optimisation results

Theorem

Let $c\colon X \nrightarrow Y$ be a channel, with prior state $\sigma \in \mathcal{D}(X)$.

(1) *Pearl increases validity: for a predicate q on Y,*

$$(c \gg= \sigma_P) \models q \;\geq\; (c \gg= \sigma) \models q \quad \text{for} \quad \sigma_P = \sigma|_{c \ll= q}.$$

(2) *Jeffrey decreases divergence: for a state τ on Y,*

$$D_{KL}(\tau, c \gg= \sigma_J) \leq D_{KL}(\tau, c \gg= \sigma) \qquad \text{for} \qquad \sigma_J = c^\dagger_\sigma \gg= \tau.$$

- The proof of Pearly is easy, but for Jeffrey it is remarkably hard.
- Jeffrey's KL-decrease is missing in the predictive coding literature — although it forms the basis of error reduction

Where we are, so far

Concluding remarks

- Updating is one of the magical things in probabilistic logic
  - it is a pillar of the AI-revolution
  - it requires a proper logic, for causality and for 'XAI'
- The two update rules of Pearl and Jeffrey:
  - can give wildly different outcomes
  - are not so clearly distinguished in the literature — probably because fuzzy / soft predicates are not standard
  - have clear formulations in terms of channels: Pearl increases validity, Jeffrey decreases divergence
- The difference Pearl / Jeffrey is of wider significance
  - e.g. EM and LDA decrease divergence via Jeffrey, see Wollic'23
- Extensions to continuous (or quantum) settings are next steps.
- Also: connecting to cognition theory community
  - hopefully this workshop gives an impetus!

iCIS | Digital Security
Radboud University