Formal Semantics of Influence in Bayesian Reasoning

Radboud University Nijmegen

Bart Jacobs, joint work with Fabio Zanasi (UCL)
bart@cs.ru.nl
Aug. 25, 2017

Outline

Where we are, so far

An exercise in Bayesian reasoning

Consider the state of a person via likelihoods of disease (D) and good mood (M).

5%	chance of disease and good mood
50%	chance of disease and no good mood
40%	chance of no disease and good mood
5%	chance of no disease and no good mood

We write this 'joint state' as a convex combination (probability distribution), using 'ket' notation $|-\rangle$

$$\sigma = 0.05|\, D, M\,\rangle + 0.5|\, D, M^{\perp}\,\rangle + 0.4|\, D^{\perp}, M\,\rangle + 0.05|\, D^{\perp}, M^{\perp}\,\rangle$$

There is a medical test for the disease, wich comes out positive in 90% of disease cases, and 5% in non-disease cases.

Question: What is the mood before and after a positive test?

before/prior: 0.45 **after/posterior:** 0.126

Which notions are required for an answer?

- states (discrete probability distributions)
  - with marginalisation for joint states, on multiple domains
- predicates, in fuzzy form, with values in $[0, 1]$
  - with a notion of validity of a predicate in a state
- conditioning (updating) a state with a predicate
- channels, for conditional probabilities (as in the test)
- comparison of a prior state and a posterior state, updated with a predicate
  - this will determine the influence of the predicate on the state

These notions will be explained first

Discrete probability distributions

Notation

- Fair coin: $\frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$
- Fair dice: $\frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$

ket notation

- $|-\rangle$ is pure syntactic sugar — stemming from quantum
- more confusing to omit them, as in: $\frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6$

- Write $\mathcal{D}(X)$ for the set of such probability distributions $\sum_i r_i|x_i\rangle$ where $x_i \in X$, $r_i \in [0, 1]$ with $\sum_i r_i = 1$
- Distributions $\omega \in \mathcal{D}(X)$ will often be called states of X

Predicates for probabilistic logic

- A predicate on a set X is a function $p\colon X \to [0, 1]$
  - It is called sharp (non-fuzzy) if $p(x) \in \{0, 1\}$ for each $x \in X$
- Basic effect module structure on these predicates:
  - true and false, as constant-one and constant-zero
  - orthosupplement $p^\perp(x) = 1 - p(x)$
  - partial sum $(p \owedge q)(x) = p(x) + q(x)$ if $p(x) + q(x) \leq 1$ for all x
  - scaling $(r \cdot p)(x) = r \cdot p(x)$, for $r \in [0, 1]$

Combining states and predicates

Let $\omega \in \mathcal{D}(X)$ be state/distribution, $p \in [0, 1]^X$ a predicate, both on X.

- **Validity** $\omega \models p$, in $[0, 1]$
  - defined as $\sum_x \omega(x) \cdot p(x)$
  - also known as expected value of p in state ω
- **Conditioning** $\omega|_p$, in $\mathcal{D}(X)$
  - assuming validity $\omega \models p$ is non-zero
  - defined as: $\omega|_p = \sum_x \frac{\omega(x) \cdot p(x)}{\omega \models p}|x\rangle$

Validity and conditioning example

- Take $X = \{1, 2, 3, 4, 5, 6\}$ with state dice $\in \mathcal{D}(X)$
  - recall dice $= \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$
- Take even predicate $E \in [0,1]^X$; it's sharp, given by:
  - $E(1) = E(3) = E(5) = 0$, $E(2) = E(4) = E(6) = 1$
  - define odd via orthosupplement: $O = E^\perp$
- dice $\models E \;=\; \frac{1}{2}$
- dice$|_E \;=\; \frac{1/6}{1/2}|2\rangle + \frac{1/6}{1/2}|4\rangle + \frac{1/6}{1/2}|6\rangle = \frac{1}{3}|2\rangle + \frac{1}{3}|4\rangle + \frac{1}{3}|6\rangle$
- dice$|_E \models O \;=\; 0$

Channels

Recall: we write $\mathcal{D}(Y)$ for the set of probability distributions on a set Y

- A channel is a function $X \to \mathcal{D}(Y)$
  - it's an X-indexed collection of distributions on Y
  - representing conditional probabilities $P(y \mid x)$
  - alternatively, it's a stochastic matrix (columns sum to 1)
  - categorically, it's a Kleisli map of the monad $\mathcal{D}$
- For a channel $c\colon X \to \mathcal{D}(Y)$ there are state- and predicate-transformations

$$\mathcal{D}(X) \xrightarrow{c_*} \mathcal{D}(Y) \qquad\qquad [0,1]^Y \xrightarrow{c^*} [0,1]^X$$

- Explicitly,

$$c_*(\omega)(y) = \textstyle\sum_x \omega(x) \cdot c(x)(y) \qquad c^*(q)(x) = \textstyle\sum_y c(x)(y) \cdot q(y).$$

- We have an equality of validities: $c_*(\omega) \models q$ is $\omega \models c^*(q)$.

Where we are, so far

What is a Bayesian network?

Short answer: a DAG in $\mathcal{K}\ell(\mathcal{D})$

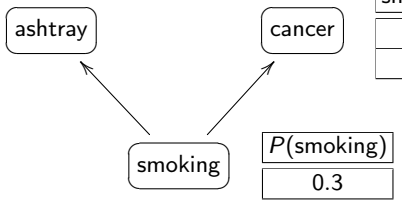
Longer: a directed graph, whose nodes with n inputs form a:

- conditional probability table, with 2^n probabilities, for all yes/no input options, or equivalently,
- a channel (Kleisli map) $2^n \to \mathcal{D}(2)$, where $2 = \{t, f\}$.

(Multiple outputs of a node are handled via copying)

Example Bayesian network I

smoking	P(ashtray)
t	0.95
f	0.25



smoking	P(cancer)
t	0.4
f	0.05

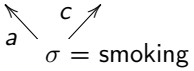
P(smoking)
0.3

- Formally, 'smoking' is a state $\sigma = 0.3|\,t\,\rangle + 0.7|\,f\,\rangle \in \mathcal{D}(2)$
- 'ashtray' and 'cancer' are channels $a, c\colon 2 \to \mathcal{D}(2)$, namely:

$$\begin{cases} a(t) = 0.95|\,t\,\rangle + 0.05|\,f\,\rangle \\ a(f) = 0.25|\,t\,\rangle + 0.75|\,f\,\rangle \end{cases} \qquad \begin{cases} c(t) = 0.4|\,t\,\rangle + 0.6|\,f\,\rangle \\ c(f) = 0.05|\,t\,\rangle + 0.95|\,f\,\rangle \end{cases}$$

Example Bayesian network II

Recall the essence of the Bayesian network structure:



We compute the:

- **cancer probability**, via state transformation c_*

$$c_*(\sigma) = 0.155|\,t\,\rangle + 0.845|\,f\,\rangle$$

- **cancer probability after observing an ashtray**, via predicate transformation a^*, conditioning $|$, and state transformation c_*

$$c_*\Big(\sigma\big|_{a^*(tt)}\Big) = 0.267|\,t\,\rangle + 0.733|\,f\,\rangle$$

Thus, the presence (or absence) of ashtrays influences cancer.

Evidence of (non-)smoking blocks the influence:

This can be expressed via formulas as: for each predicate p,

$$c_*\Big(\sigma\big|_{tt}\Big) = c_*\Big(\sigma\big|_{a^*(p)\&tt}\Big) \qquad c_*\Big(\sigma\big|_{ff}\Big) = c_*\Big(\sigma\big|_{a^*(p)\&ff}\Big)$$

This is an example of d-separation in Bayesian network theory.

Three forms of d-separation

(i) **fork** connection $\boxed{A} \leftarrow \boxed{B} \rightarrow \boxed{C}$, information on A will influence C and viceversa, but this flow is blocked once B is known

(ii) **serial** $\boxed{A} \rightarrow \boxed{B} \rightarrow \boxed{C}$, event A influences C through B (and viceversa), but knowledge of B blocks this mutual influence

(iii) **collider** situation $\boxed{A} \rightarrow \boxed{B} \leftarrow \boxed{C}$, any evidence about B (and its descendants) will make A and C depend on each other.

Message of the paper

- These influences & blocks can be precisely explained via state- and predicate- transformation — after viewing a Bayesian network as a graph of Kleisli maps
- This blocking only works for sharp $\{0, 1\}$-valued evidence — not for fuzzy $[0, 1]$-valued predicates

(Bayesian network books all use "handwaving" explanations)

Where we are, so far

Basic questions, that started the work

- How different are a state ω and and an update $\omega|_p$?
- Can this be formalised as distance $d(\omega, \omega|_p)$?
  - such a number would capture the influence of p on ω
- Question that emerged along the way: what about joint states, and updating in one component only?
  - like in the disease-mood example

Distances in probability theory

- Much usage of metrics in probabilistic computation, eg.
  - measuring the behavioural similarity of states in probabilistic transition systems
    (Panangaden, Desharnais, Mislove, Worrell, van Breughel, König ...)
  - evaluating performance and uncertainty of Bayesian network models
- Nice systematic description of metrics via universality in LICS'16 paper of Madare, Panangaden, Plotkin
- Here, the metric is a parameter; its choice is not essential
- The metric is used to measure the influence of a predicate on a state, via conditioning
  - direct influence, for matching states & predicates
  - crossover influence for joint states, with predicate acting on one component

Total variation distance

The 'total variation' distance on $\omega, \sigma \in \mathcal{D}(X)$ is:

$$d(\omega, \sigma) = \tfrac{1}{2} \sum_x |\omega(x) - \sigma(x)|$$

Some remarks

- this is special (discrete) case of Kantorovic metric, defined for distributions on metric spaces
- it takes values in $[0, 1]$
- it can be alternatively described via 'couplings', and also via predicates and validity as:

$$d(\omega, \sigma) = \bigvee_{p \in [0,1]^X} |\, \omega \models p - \sigma \models p \,|$$

Direct influence

Let $\omega \in \mathcal{D}(X)$ and $p \in [0, 1]^X$.

Definition

The (direct) influence of the predicate p on the state ω is the distance:

$$d(\omega, \omega|_p)$$

Example

Take $2 = \{t, f\}$ with predicate $p(t) = 0.9, p(f) = 0.05$

$$d(\omega, \omega|_p) = 0.19 \quad \text{for } \omega = 0.8|\, t \,\rangle + 0.2|\, f \,\rangle$$
$$d(\sigma, \sigma|_p) = 0.45 \quad \text{for } \sigma = 0.5|\, t \,\rangle + 0.5|\, f \,\rangle$$
$$d(\tau, \tau|_p) = 0.62 \quad \text{for } \tau = 0.2|\, t \,\rangle + 0.8|\, f \,\rangle$$

One can prove:

$$d(\omega, \omega|_p) = 0 \iff \omega = \omega|_p \iff p \text{ is contant \& non-zero}$$

Products and marginalisations of states

For another form of 'crossover' influence we need joint states.

- For states $\omega_1 \in \mathcal{D}(X_1)$ and $\omega_2 \in \mathcal{D}(X_2)$ we can form the product state $\omega_1 \otimes \omega_2 \in \mathcal{D}(X_1 \times X_2)$ by:

$$(\omega_1 \otimes \omega_2)(x_1, x_2) = \omega_1(x_1) \cdot \omega_2(x_2)$$

- For a joint state $\sigma \in \mathcal{D}(X_1 \otimes X_2)$ there are marginalisations $M_i(\sigma) \in \mathcal{D}(X_i)$, given by:

$$M_1(\sigma)(x_1) = \textstyle\sum_{x_2} \sigma(x_1, x_2) \quad M_2(\sigma)(x_2) = \sum_{x_1} \sigma(x_1, x_2)$$

- It is to easy that marginalisation after product returns the originals:

$$M_1(\omega_1 \otimes \omega_2) = \omega_1 \qquad M_2(\omega_1 \otimes \omega_2) = \omega_2$$

- But in general: $\sigma \neq M_1(\sigma) \otimes M_2(\sigma)$

Entwinedness of joint states

Take $\sigma = \frac{1}{2}|\, a, 1\,\rangle + \frac{1}{2}|\, b, 2\,\rangle$

- $M_1(\sigma) = \frac{1}{2}|\, a\,\rangle + \frac{1}{2}|\, b\,\rangle$, $M_2(\sigma) = \frac{1}{2}|\, 1\,\rangle + \frac{1}{2}|\, 2\,\rangle$
- And: $M_1(\sigma) \otimes M_2(\sigma) = \frac{1}{4}|\, a, 1\,\rangle + \frac{1}{4}|\, a, 2\,\rangle + \frac{1}{4}|\, b, 1\,\rangle + \frac{1}{4}|\, b, 2\,\rangle$
- Thus, indeed: $\sigma \neq M_1(\sigma) \otimes M_2(\sigma)$

This does not happen for cartesian products; it is typical for tensors — of the probabilistic kind, with projections

There are various expressions for this phenomenon: the components of the joint states are dependent, correlated, entangled (quantum), entwined

Another manifestation of entwinedness

Let $\sigma \in \mathcal{D}(X \times Y)$ and $p \in [0,1]^X$

- We can 'weaken' p to $p \otimes \mathbf{1} \in [0,1]^{X \times Y}$, with $(p \otimes \mathbf{1})(x, y) = p(x)$.
- The domains now match, so we can update σ to $\sigma|_{p \otimes \mathbf{1}}$
- Predicate p works on the X component. Do we see any difference in the Y component after updating?

Answer: Yes !!

More precisely: $M_2(\sigma) \neq M_2(\sigma|_{p \otimes \mathbf{1}})$ in general, not if σ is a product

Definition

The crossover influence of p on σ is:

$$d\big(M_2(\sigma),\ M_2(\sigma|_{p \otimes \mathbf{1}})\big)$$

We can now solve the disease-mood question!

Recall: $\sigma = 0.05|\, D, M\,\rangle + 0.5|\, D, M^\perp\,\rangle + 0.4|\, D^\perp, M\,\rangle + 0.05|\, D^\perp, M^\perp\,\rangle$

- $M_2(\sigma) = 0.45|\, M\,\rangle + 0.55|\, M^\perp\,\rangle$
- test predicate $p(D) = 0.9$, $p(D^\perp) = 0.05$
- weakened test: $(p \otimes \mathbf{1})(D, M) = (p \otimes \mathbf{1})(D, M^\perp) = 0.9$ and $(p \otimes \mathbf{1})(D^\perp, M) = (p \otimes \mathbf{1})(D^\perp, M^\perp) = 0.05$
- $\sigma \models p \otimes \mathbf{1} = 0.05 \cdot 0.9 + 0.5 \cdot 0.9 + 0.4 \cdot 0.05 + 0.05 \cdot 0.05 = 0.5175$
- $\sigma|_{p \otimes \mathbf{1}}$
  $= \frac{0.05 \cdot 0.9}{0.5175}|\, D,M\,\rangle + \frac{0.5 \cdot 0.9}{0.5175}|\, D,M^\perp\,\rangle + \frac{0.4 \cdot 0.05}{0.5175}|\, D^\perp,M\,\rangle + \frac{0.05 \cdot 0.05}{0.5175}|\, D^\perp,M^\perp\,\rangle$
  $= 0.087|\, D,M\,\rangle + 0.869|\, D,M^\perp\,\rangle + 0.039|\, D^\perp,M\,\rangle + 0.005|\, D^\perp,M^\perp\,\rangle$
- Thus: $M_2(\sigma|_{p \otimes \mathbf{1}}) = 0.126|\, M\,\rangle + 0.874|\, M^\perp\,\rangle$.
- A clear example of crossover influence.

Where we are, so far

Final remarks

- Powerful perspective: Bayesian network is:
  - directed acyclic graph of channels
  - i.e. DAG in $\mathcal{K}\ell(\mathcal{D})$
- Also powerful use of state/predicate-transformation and updating
  - **forward inference**: first update, then state-transformation
  - **backward inference**: first predicate-transformation, then update

  Introduced in (Jacobs-Zanasi, MFPS'16)
- Two main contributions here:
  (1) Explanation of d-separation in Bayesian networks using these terms
  (2) Metric approach to influence, also in 'crossover' form
- **Aside:** We have a tool EfProb in which one can compute with states, predicates, updates, channels etc, for discrete, continuous, and quantum probability, see `efprob.cs.ru.nl`

iCIS | Digital Security
Radboud University