Learning along a Channel: the Expectation part of Expectation-Maximisation

Radboud University Nijmegen
MFPS, June 6, 2019

Bart Jacobs
bart@cs.ru.nl

Outline

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

Where we are, so far

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

Setting and topic

- Ever since Lawvere & Giry in the early 1980s, we know that there is much (categorical) structure in probability
  - a monads of distributions, both continuous and discrete: $\mathcal{G}$ and $\mathcal{D}$
  - their Kleisli categories are models of computation
  - these monads are commutative/monoidal and affine and ...
- Since then, the area has been rather silent
- There is a recent revival, with the grown interest in probabilistic programming
  - much work on higher order probabilistic models
  - but also on sampling and conditioning
  - Bayesian reasoning in Kleisli categories
  - this work dives into probabilistic learning — of parameters, not of graph structure

Distributions (states) & predicates, discretely

A (discrete probability) distribution is a formal convex combination:

$$\omega = \tfrac{1}{3}|\,a\,\rangle + \tfrac{1}{2}|\,b\,\rangle + \tfrac{1}{6}|\,c\,\rangle \qquad \text{on} \qquad X = \{a, b, c, \ldots\}$$

This ω is a function $X \to [0,1]$ with values adding up to 1.

- we write $\mathcal{D}(X)$ for such distributions on X; this gives a monad.

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

- We write $Pred(X)$ for the set of predicates on X; it is an effect module
- Each subset/event $E \subseteq X$ forms a 'sharp' predicate, via the indicator function $\mathbf{1}_E\colon X \to [0,1]$
- One can also work with factors $p\colon X \to \mathbb{R}_{\geq 0}$, which form a commutative monoid

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.

Link with traditional notation for $E, D \subseteq X$, and ω implicit

$$P(E) \;=\; \omega \models \mathbf{1}_E \qquad \text{and} \qquad P(D \mid E) \;=\; \omega|_{\mathbf{1}_E} \models \mathbf{1}_D.$$

Learning in basic form (own interpretation)

$$\omega \quad \models \quad p$$

distribution/state

predicate/evidence

- Learning is about changing one's state ω in order to increase the validity: it's about getting a better match with the evidence p.
- Learning algorithms do this iteratively, via each time turning ω into ω' so that $\omega' \models p \;\geq\; \omega \models p$

Theorem (1)

$$\omega \models p \;\;\leq\;\; \omega|_p \models p$$

This is intuitively clear, but not easy to prove (it's not in the MFPS-paper)

Intermezzo on state & predicate transformation

A channel $c\colon X \rightarrow Y$ is a Kleisli map $c\colon X \to \mathcal{D}(Y)$.

(1) It turns a state $\omega \in \mathcal{D}(X)$ into a state $c \gg \omega \in \mathcal{D}(Y)$ via:

$$c \gg \omega \;:=\; \sum_y \big(\textstyle\sum_x \omega(x) \cdot c(x)(y)\big)\,|\,y\,\rangle.$$

(2) It turns a predicate $q \in [0,1]^Y$ into a predicate $c \ll q \in [0,1]^X$, where:

$$(c \ll q)(x) \;:=\; \sum_y c(x)(y) \cdot q(y).$$

Lemma

$$c \gg \omega \models q \;=\; \omega \models c \ll q$$

Where we are, so far

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

A coin with observations

Assume I have a fair coin $\sigma = \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$.

(1) What is the likelihood of getting two heads?

(2) What is the likelihood of getting one head, one tail?

(3) What is the likelihood of the predicates p, q with:

$$\begin{cases} p(H) = 0.8 \\ p(T) = 0.2 \end{cases} \qquad \begin{cases} q(H) = 0.6 \\ q(T) = 0.4 \end{cases}$$

In all these cases there are two possible answers, depending on whether one uses the coin once (with two observers) or twice (with one observer).

- this is similar to draws from an urn with or without replacement

A more systematic approach via products

For states $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ there is $\omega \otimes \rho \in \mathcal{D}(X \times Y)$ via:

$$\omega \otimes \rho \coloneqq \sum_{x,y} \omega(x) \cdot \rho(y) \big| x, y \big\rangle.$$

For predicates there are two products/conjunctions & and $\otimes$

(1) the parallel conjunction: for $p \in [0,1]^X$ and $q \in [0,1]^Y$

$$X \times Y \xrightarrow{p \otimes q} [0,1] \quad \text{given by} \quad (x,y) \longmapsto p(x) \cdot q(y).$$

(2) the sequential conjunction: for $p_1, p_2 \in [0,1]^X$ on the same set:

$$X \xrightarrow{p_1 \,\&\, p_2} [0,1] \quad \text{given by} \quad x \longmapsto p_1(x) \cdot p_2(x).$$

Products and validity

For parallel conjunction $\otimes$ we have:

Lemma

$$\omega \otimes \rho \models p \otimes q \;=\; (\omega \models p) \cdot (\rho \models q)$$

For sequential conjunction & we have:

Lemma

$$\omega \models p_1 \,\&\, p_2 \;\neq\; (\omega \models p_1) \cdot (\omega \models p_2)$$

But we do have:

$$\omega \models p_1 \,\&\, p_2 \;=\; \omega \models \Delta \ll (p_1 \otimes p_2) \;=\; \Delta \gg \omega \models p_1 \otimes p_2.$$

Important difference: $\begin{cases} \text{multiple state perspective} & \omega \otimes \omega \\ & \not\models \\ \text{copied state perspective} & \Delta \gg \omega \end{cases}$

Coin with observations, revisited

We use a fair coin state $\sigma = \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$.

(1) What is the likelihood of getting two heads?
M: $\sigma \otimes \sigma \models \mathbf{1}_H \otimes \mathbf{1}_H \;=\; (\sigma \models \mathbf{1}_H) \cdot (\sigma \models \mathbf{1}_H) \;=\; \frac{1}{4}$
C: $\sigma \models \mathbf{1}_H \,\&\, \mathbf{1}_H \;=\; \sigma \models \mathbf{1}_H \;=\; \frac{1}{2}$

(2) What is the likelihood of getting one head, one tail?
M: $\sigma \otimes \sigma \models \mathbf{1}_H \otimes \mathbf{1}_T \;=\; (\sigma \models \mathbf{1}_H) \cdot (\sigma \models \mathbf{1}_T) \;=\; \frac{1}{4}$
C: $\sigma \models \mathbf{1}_H \,\&\, \mathbf{1}_T \;=\; \sigma \models \mathbf{0} \;=\; 0$

(3) What is the likelihood of $p = 0.8 \cdot \mathbf{1}_H + 0.2 \cdot \mathbf{1}_T$ and $p = 0.6 \cdot \mathbf{1}_H + 0.4 \cdot \mathbf{1}_T$?
M: $\sigma \otimes \sigma \models p \otimes q \;=\; (\sigma \models p) \cdot (\sigma \models q) \;=\; \frac{1}{4}$
C: $\sigma \models p \,\&\, q \;=\; \sigma \models 0.48 \cdot \mathbf{1}_H + 0.08 \cdot \mathbf{1}_T \;=\; 0.28$

Where we are, so far

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

What is data?

- Data for learning typically comes in sequences or tables. The order does not matter (in updating), but multiple occurrences of the same items are relevant.
- Hence we use multisets for data
- There is a monad for this, written as $\mathcal{M}$, where:

$$\mathcal{M}(X) \coloneqq \{\varphi\colon X \to \mathbb{N} \mid supp(\varphi) \text{ is finite}\}$$

There are two representations of data on X:

(1) pointwise: simply use $\mathcal{M}(X)$

(2) predicate-wise: use $\mathcal{M}(Pred(X))$

Representation (2) is new, but makes much sense if we wish to deal with uncertainties about data; it subsumes (1) via point predicates $\mathbf{1}_x$.

Validity of data

- Suppose we have a state $\omega \in \mathcal{D}(X)$ and data $\Phi \in \mathcal{M}(Pred(X))$
- What is the validity of Φ in ω?
- It is this validity that we wish to increase in learning

(1) Multiple state interpretation

$$\omega \models_{\mathsf{M}} \Phi \;\coloneqq\; \prod_p \,(\omega \models p)^{\Phi(p)}$$

(2) Copied state interpretation

$$\omega \models_{\mathsf{C}} \Phi \;\coloneqq\; \omega \models \&_p \; p^{\Phi(p)}$$

- There are thus also two forms of learning, for $\models_{\mathsf{M}}$ and for $\models_{\mathsf{C}}$
- I have not seen this distinction in the literature . . .

Basic result for M-learning

Theorem (2)

$$\omega \models_{M} \Phi \;\leq\; \omega' \models_{M} \Phi$$

for:

$$\omega' \;\coloneqq\; \sum_{p} \frac{\Phi(p)}{|\Phi|} \cdot \omega|_{p} \qquad \textit{where} \qquad |\Phi| \;\coloneqq\; \textstyle\sum_{p} \Phi(p).$$

- Proof is not easy, result is not in the paper
- When $\Phi = \sum_{x} \Phi(x) | x \rangle$ is pointwise data, i.e. $\Phi \in \mathcal{M}(X)$, we get normalisation of the multiset:

  $$Flrn(\Phi) \;\coloneqq\; \omega' \;=\; \textstyle\sum_{x} \frac{\Phi(x)}{|\Phi|} | x \rangle$$

  where $Flrn$ stands for frequentist learning (by counting)
- C-learning can be done via Theorem 1: $\omega \models_{C} \Phi \;\leq\; \omega|_{\&_{p} p^{\Phi(p)}} \models_{C} \Phi$

Results about frequentist learning (in the paper)

Theorem

Frequentist learning is a natural transformation:

$$Flrn \colon \mathcal{M}_{*} \Longrightarrow \mathcal{D}$$

It is monoidal and commutes with extraction (disintegration)

Theorem (classical)

For $\varphi \in \mathcal{M}(X)$, the function:

$$\mathcal{D}(X) \xrightarrow{\;(-) \models_{M} \varphi\;} [0,1]$$

reaches its maximum at $Flrn(\varphi)$. Hence $\omega \models_{M} \varphi \;\leq\; Flrn(\varphi) \models_{M} \varphi$.

Where we are, so far

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

What is Expectation-Maximisation (EM)?

- It is an iterative algorithm for learning
  - due to: Arthur Dempster, Nan Laird, and Donald Rubin (1977)
  - widely-used in many situations, also for Markov chains / HMMs
- The term "EM" has developed into an umbrella term
  - It is applied differently in different situations; what EM is in general is unclear (to me)
- The paper elaborates two examples, with different EM-interpretations:
  - from classic book: Russell-Norvig, *Artifical Intelligence*
  - from influential article: Do & Batzoglou, *What is the expectation maximization algorithm?* in Nature Biotechnology, 2008.
- The difference can be explained in terms of M-learning versus C-learning

EM-essentials: state-and-channel learning

- We considered situations with state and data on the same set X
- But frequently we like to learn about a set X whereas we have data on a different set Y
  - typically this happens in classification or clustering

$$1 \xrightarrow{\omega} X \xrightarrow{c} Y$$

learning aim (X) data (Y)

- In EM we like to learn both:
  - the E-part: a state $\omega \in \mathcal{D}(X)$, i.e. $\omega \colon 1 \to X$
  - the M-part: a channel $c \colon X \to Y$
- Here, and in the paper, we concentrate on the state (E-part)
  - Concretely: given a state ω and channel c, we aim to learn a "better" ω' — and also c'

The candy example, from Russell-Norvig, §20.3

We consider a bag with two types of candies (0 and 1), which can have:

- two flavours, cherry (C) or lime (L)
- a red (R) or green (G) wrapper
- a hole (H) or not ($H^\perp$)

These probabilities of these properties for each sort of candies are given by three channels, written as

$$f \colon \{0,1\} \to \{C, L\} \quad w \colon \{0,1\} \to \{R, G\} \quad h \colon \{0,1\} \to \{H, H^\perp\}$$

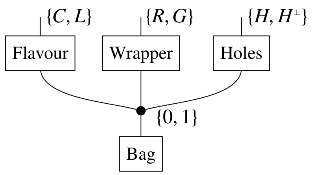
with:

$$\begin{array}{rclrcl} f(0) & = & \frac{6}{10}|\,C\,\rangle + \frac{4}{10}|\,L\,\rangle & f(1) & = & \frac{4}{10}|\,C\,\rangle + \frac{6}{10}|\,L\,\rangle \\ w(0) & = & \frac{6}{10}|\,R\,\rangle + \frac{4}{10}|\,G\,\rangle & w(1) & = & \frac{4}{10}|\,R\,\rangle + \frac{6}{10}|\,G\,\rangle \\ h(0) & = & \frac{6}{10}|\,H\,\rangle + \frac{4}{10}|\,H^\perp\,\rangle & h(1) & = & \frac{4}{10}|\,H\,\rangle + \frac{6}{10}|\,H^\perp\,\rangle \end{array}$$

The initial candy distribution is: $\rho = \frac{6}{10}|\,0\,\rangle + \frac{4}{10}|\,1\,\rangle$

Candy example, part II: the data

We thus have a Bayesian network (as string diagram):



The data to learn from is a multiset $\psi \in \mathcal{M}(\{C, L\} \times \{R, G\} \times \{H, H^\perp\})$

$$\begin{aligned} \psi = {} & 273|\,C, R, H\,\rangle + 93|\,C, R, H^\perp\,\rangle + 104|\,C, G, H\,\rangle + 90|\,C, G, H^\perp\,\rangle \\ & + 79|\,L, R, H\,\rangle + 100|\,L, R, H^\perp\,\rangle + 94|\,L, R, H\,\rangle + 167|\,L, R, H^\perp\,\rangle. \end{aligned}$$

How to learn a new candy-in-the-bag distribution ρ' on $\{0, 1\}$?

Candy example, part III: the analysis

- We combine the three channels into a 3-tuple:

$$\{0,1\} \xrightarrow{\langle f, w, h\rangle} \{C, L\} \times \{R, G\} \times \{H, H^\perp\}$$

- We wish to increase the M-validity:

$$\begin{aligned} \langle f, w, h\rangle \gg \rho \models_{\mathsf{M}} \psi & = \prod_d \Big(\langle f, w, h\rangle \gg \rho \models \mathbf{1}_d\Big)^{\psi(d)} \\ & = \prod_d \Big(\rho \models \langle f, w, h\rangle \ll \mathbf{1}_d\Big)^{\psi(d)} \end{aligned}$$

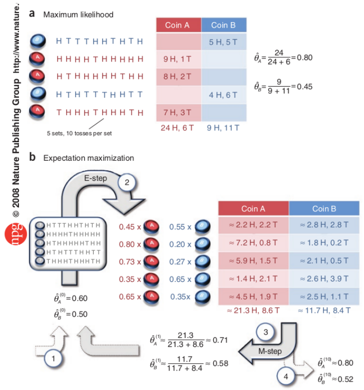
- Theorem 2 gives a formula for a better state ρ', with increased validity:

$$\rho' = \sum_d \frac{\psi(d)}{|\psi|} \cdot \rho\big|_{\langle f, w, h\rangle \ll \mathbf{1}_d}$$

- The outcome is exactly as given in Russell-Norvig
  - but there, only a formula is given that is claimed to be EM, without explanation or proof
  - our account can also be described as "dagger" of a channel

Coin example, from Do & Batzoglou 2008

Explanation by example, via a often-reproduced picture, for applications in gene expression clustering in computational biology:



Coin example, part II: channel-based analysis

- We have two coins (0 and 1), each with their own bias; the aim is to learn both the distribution of coins and the associated biases from data
- There is a given channel c and state ω in:

$$\{0,1\} \xrightarrow{c} \{H,T\} \qquad \text{with} \qquad \omega \in \mathcal{D}(\{0,1\})$$

- Learning starts from the uniform state $\omega = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle$ with channel:

$$c(0) = \tfrac{3}{5}|H\rangle + \tfrac{2}{5}|T\rangle \qquad \text{and} \quad c(1) = \tfrac{1}{2}|H\rangle + \tfrac{1}{2}|T\rangle.$$

- The aim is to find better ω' and c'. We concentrate on ω'.

Coin example, part III: analysis

- The data are given in the form of a multiset $\psi \in \mathcal{M}(\{H,T\})$ of heads and tails
- The Do-Batzoglou example uses C-learning, via validity:

$$\omega \models_{\mathcal{C}} \&_d \, (c \ll \mathbf{1}_d)^{\psi(d)}$$

- A better state ω' is obtained via conditioning (Theorem 1):

$$\omega' := \omega\big|_{\&_d (c \ll \mathbf{1}_d)^{\psi(d)}}$$

- This gives precisely the outcomes of Do-Batzoglou.

Brief comparison of M-learning and C-learning

Using the coin data $\psi_1, \ldots, \psi_5 \in \mathcal{M}(\{H,T\})$ of Do-Batzoglou we get:

data ψ_i	**C-learning**	**M-learning**
$5|H\rangle + 5|T\rangle$	$0.4491|0\rangle + 0.5509|1\rangle$	$0.4949|0\rangle + 0.5051|1\rangle$
$9|H\rangle + 1|T\rangle$	$0.805|0\rangle + 0.195|1\rangle$	$0.5354|0\rangle + 0.4646|1\rangle$
$8|H\rangle + 2|T\rangle$	$0.7335|0\rangle + 0.2665|1\rangle$	$0.5253|0\rangle + 0.4747|1\rangle$
$4|H\rangle + 6|T\rangle$	$0.3522|0\rangle + 0.6478|1\rangle$	$0.4848|0\rangle + 0.5152|1\rangle$
$7|H\rangle + 3|T\rangle$	$0.6472|0\rangle + 0.3528|1\rangle$	$0.5152|0\rangle + 0.4848|1\rangle$

It seems that C-learning is better at picking up the differences.

Where we are, so far

Introduction

Multiple-state and copied-state perspectives

Data, as input for learning

Expectation-Maximisation

Conclusions

Concluding remarks

- Probabilistic learning is a fascinating topic, of great relevance today, in probabilistic data analysis and AI
- Proposed definition of learning: increasing the validity of data, via "better" state (and channel)
- There is lots of (categorical) structure, which is traditionally left implicit
- There are also fundamentally distinct perspectives:
  - multiple state: $\models_{M}$ and M-learning
  - copied state: $\models_{C}$ and C-learning

  Again, these distinctions are left implicit.
- Versions of EM in the literature can be explained via $\models_{M}$ and $\models_{C}$
  - We've shown how to get 'better' states, not 'better' channels
- Many details of this talk are still unpublished, also about Baum-Welch for hidden Markov models.