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

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: $G$ and $D$
- 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:

\[ \omega = \frac{1}{3} |a \rangle + \frac{1}{2} |b \rangle + \frac{1}{6} |c \rangle \text{ on } X = \{a, b, c, \ldots \} \]

This \( \omega \) is a function \( X \rightarrow [0, 1] \) with values adding up to 1.

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

A predicate on a set \( X \) is an arbitrary function \( p : X \rightarrow [0, 1] \).

- We write \( \text{Pred}(X) \) 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 \( 1_E : X \rightarrow [0, 1] \)

- One can also work with factors \( p : X \rightarrow \mathbb{R}_{\geq 0} \), which form a commutative monoid.

Validity and conditioning

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

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

It describes the expected value of \( p \) in \( \omega \).

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

\[ \omega \models p(x) := \frac{\omega(x) \cdot p(x)}{\omega \models p} \]

that is \( \omega \models p = \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \cdot |x \rangle \).

It’s the normalised product of \( \omega \) and \( p \).

Link with traditional notation for \( E, D \subseteq X \), and \( \omega \) implicit

\[ P(E) = \omega \models 1_E \text{ and } P(D \mid E) = \omega \models 1_D \]

Learning in basic form (own interpretation)

\[ \omega \models p \]

distribution/state \hspace{1cm} predicate/evidence

- Learning is about changing one’s state \( \omega \) 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 \( \omega \) into \( \omega' \) so that \( \omega' \models p \geq \omega \models p \)

Theorem (1)

\[ \omega \models p \leq \omega \models p \]

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

Intermezzo on state & predicate transformation

A channel \( c : X \rightarrow Y \) is a Kleisli map \( c : X \rightarrow D(Y) \).

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

\[ c \gg \omega := \sum_y \langle \sum_x \omega(x) \cdot c(x)(y) \mid y \rangle. \]

(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

\[ c \gg \omega \models q = \omega \models c \ll q \]
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:
   \[
   \begin{align*}
   p(H) &= 0.8 \\
   p(T) &= 0.2 \\
   q(H) &= 0.6 \\
   q(T) &= 0.4
   \end{align*}
   \]

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 \( \omega \in D(X) \) and \( \rho \in D(Y) \) there is \( \omega \otimes \rho \in D(X \times Y) \) via:

\[
\omega \otimes \rho := \sum_{x,y} \omega(x) \cdot \rho(y) | x, y \rangle.
\]

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

1. the parallel conjunction: for \( p \in [0,1]^X \) and \( q \in [0,1]^Y \)
   \[
   X \times Y \xrightarrow{p \otimes q} [0,1] \quad \text{given by} \quad (x,y) \mapsto p(x) \cdot q(y).
   \]

2. the sequential conjunction: for \( p_1, p_2 \in [0,1]^X \) on the same set:
   \[
   X \xrightarrow{p_1 \otimes p_2} [0,1] \quad \text{given by} \quad x \mapsto p_1(x) \cdot p_2(x).
   \]

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: \[
\begin{align*}
\text{multiple state perspective} & \quad \omega \otimes \omega \\
\text{copied state perspective} & \quad \Delta \gg \omega
\end{align*}
\]
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:** \( \sigma \otimes \sigma \models 1_H \otimes 1_H = (\sigma \models 1_H) \cdot (\sigma \models 1_H) = \frac{1}{4} \)
   - **C:** \( \sigma \models 1_H \& 1_H = \sigma \models 1_H = \frac{1}{2} \)

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

3. What is the likelihood of \( p = 0.8 \cdot 1_H + 0.2 \cdot 1_T \) and \( p = 0.6 \cdot 1_H + 0.4 \cdot 1_T \)?
   - **M:** \( \sigma \otimes \sigma \models p \otimes q = (\sigma \models p) \cdot (\sigma \models q) = \frac{1}{2} \)
   - **C:** \( \sigma \models p \& q = \sigma \models 0.48 \cdot 1_H + 0.08 \cdot 1_T = 0.28 \)

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:
  \[
  \mathcal{M}(X) := \{ \varphi : X \to \mathbb{N} \mid \text{supp}(\varphi) \text{ is finite} \}
  \]

There are two representations of data on \( X \):

1. **Pointwise:** simply use \( \mathcal{M}(X) \)
2. **Predicate-wise:** use \( \mathcal{M}(\text{Pred}(X)) \)

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

Validity of data

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

1. **Multiple state interpretation**
   \[
   \omega \models^m \Phi := \prod_p (\omega \models p)^{\Phi(p)}
   \]

2. **Copied state interpretation**
   \[
   \omega \models^c \Phi := \omega \models p^{\Phi(p)}
   \]

- There are thus also two forms of learning, for \( \models^m \) and for \( \models^c \).
- I have not seen this distinction in the literature . . .
Basic result for M-learning

**Theorem (2)**

$$\omega \mid_{\mathcal{M}} \Phi \leq \omega' \mid_{\mathcal{M}} \Phi$$

for:

$$\omega' := \sum_p \frac{\Phi(p)}{|\Phi|} \cdot \omega|_p$$

where $$|\Phi| := \sum_p \Phi(p).$$

Proof is not easy, result is not in the paper

When $$\Phi = \sum_x \Phi(x) \mid x$$ is pointwise data, i.e. $$\Phi \in \mathcal{M}(X),$$ we get normalisation of the multiset:

$$\text{Flrn}(\Phi) := \omega' = \sum_x \frac{\Phi(x)}{|\Phi|} \mid x$$

where Flrn stands for frequentist learning (by counting)

C-learning can be done via Theorem 1: $$\omega \mid_{\mathcal{C}} \Phi \leq \omega|_{\mathcal{C},\text{prop}} \mid_{\mathcal{C}} \Phi$$

Results about frequentist learning (in the paper)

**Theorem**

*Frequentist learning is a natural transformation:*

$$\text{Flrn} : \mathcal{M}_* \Rightarrow \mathcal{D}$$

It is monoidal and commutes with extraction (disintegration)

**Theorem (classical)**

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

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

reaches its maximum at Flrn($$\varphi$$). Hence $$\omega \mid_{\mathcal{M}} \varphi \leq \text{Flrn}(\varphi) \mid_{\mathcal{M}} \varphi.$$
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

\[
\begin{array}{ccc}
\text{learning aim} & \xrightarrow{\psi} & \text{data} \\
1 \xrightarrow{\omega} X & \xrightarrow{c} & Y
\end{array}
\]

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

Candy example, part II: the data

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

- Flavour $\{C, L\}$
- Wrapper $\{R, G\}$
- Holes $\{H, H^+\}$

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

\[
\psi = 273\ | \ C, R, H \rangle + 93\ | \ C, R, H^+ \rangle + 104\ | \ C, G, H \rangle + 90\ | \ C, G, H^+ \rangle \\
+ 79\ | \ L, R, H \rangle + 100\ | \ L, R, H^+ \rangle + 94\ | \ L, R, H^+ \rangle + 167\ | \ L, R, H^+ \rangle.
\]

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

Candy example, part III: the analysis

- We combine the three channels into a 3-tuple:
  \[
  \{0, 1\} \xrightarrow{\langle f, w, h \rangle} \{C, L\} \times \{R, G\} \times \{H, H^+\}
  \]
- We wish to increase the M-validity:
  \[
  \langle f, w, h \rangle \gg \rho \quad \implies \quad \psi = \prod_d \langle f, w, h \gg \rho = 1_d \rangle^{\psi(d)} = \prod_d \langle f, w, h \gg \rho = 1_d \rangle^{\psi(d)}
  \]
- Theorem 2 gives a formula for a better state $\rho'$, with increased validity:
  \[
  \rho' = \sum_d \frac{\psi(d)}{\psi} | f, w, h \gg \rho = 1_d \rangle
  \]
- The outcome is exactly as given in Russell-Norfïg
  - 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 $\omega$ in:
  \[ \{0, 1\} \xrightarrow{c} \{H, T\} \quad \text{with} \quad \omega \in D(\{0, 1\}) \]
- Learning starts from the uniform state $\omega = \frac{1}{2} \ket{0} + \frac{1}{2} \ket{1}$ with channel:
  \[ c(0) = \frac{3}{5} \ket{H} + \frac{2}{5} \ket{T} \quad \text{and} \quad c(1) = \frac{1}{2} \ket{H} + \frac{1}{2} \ket{T}. \]
- The aim is to find better $\omega'$ and $c'$. We concentrate on $\omega'$.

**Coin example, part III: analysis**

- The data are given in the form of a multiset $\psi \in M(\{H, T\})$ of heads and tails.
- The Do-Batzoglou example uses C-learning, via validity:
  \[ \omega |\!\!| \&_{d} (c \ll 1_{d})^{(d)} \]
- A better state $\omega'$ is obtained via conditioning (Theorem 1):
  \[ \omega' := \omega |_{k_{d}(c \ll 1_{d})^{(d)}} \]
- This gives precisely the outcomes of Do-Batzoglou.

**Brief comparison of M-learning and C-learning**

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

<table>
<thead>
<tr>
<th>data $\psi_i$</th>
<th>C-learning</th>
<th>M-learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$5 \ket{H} + 5 \ket{T}$</td>
<td>0.4491\ket{0} + 0.5509\ket{1}</td>
<td>0.4949\ket{0} + 0.5051\ket{1}</td>
</tr>
<tr>
<td>$9 \ket{H} + 1 \ket{T}$</td>
<td>0.805\ket{0} + 0.195\ket{1}</td>
<td>0.5354\ket{0} + 0.4646\ket{1}</td>
</tr>
<tr>
<td>$8 \ket{H} + 2 \ket{T}$</td>
<td>0.7335\ket{0} + 0.2665\ket{1}</td>
<td>0.5253\ket{0} + 0.4747\ket{1}</td>
</tr>
<tr>
<td>$4 \ket{H} + 6 \ket{T}$</td>
<td>0.3522\ket{0} + 0.6478\ket{1}</td>
<td>0.4848\ket{0} + 0.5152\ket{1}</td>
</tr>
<tr>
<td>$7 \ket{H} + 3 \ket{T}$</td>
<td>0.6472\ket{0} + 0.3528\ket{1}</td>
<td>0.5152\ket{0} + 0.4848\ket{1}</td>
</tr>
</tbody>
</table>

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: $\|_M$ and M-learning
  - copied state: $\|_C$ and C-learning
    Again, these distinctions are left implicit.
- Versions of EM in the literature can be explained via $\|_M$ and $\|_C$
  - 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.