A principled approach to Expectation Maximisation and Latent Dirichlet Allocation using Jeffrey's update rule

Radboud University Nijmegen
Wollic, Halifax, July. 12, 2023

Bart Jacobs

bart@cs.ru.nl

Outline

Where we are, so far

Long introduction to probabilistic learning

Mathematical background

Expectation Maximisation (EM)

Conclusions

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"

My own (logical) interests/work

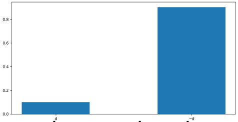
- There are two update rules, by Judea Pearl (1936) and by Richard Jeffrey (1926 – 2002)
  - They both have clear formulations using channels — see later
  - What are the differences? When to use which rule?
- Intriguing question: does the human mind use Pearl's or Jeffrey's rule — within predictive coding theory
  - cognitive science may provide an answer
- **Here:** what about machine learning algorithms, like Expectation-Maximisation (EM) and Latent Dirichlet Allocation (LDA)?

- 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

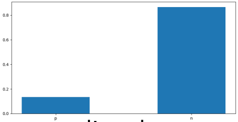
Example, medical test, part I

- 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?

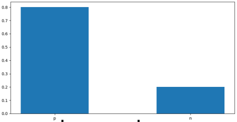
Example, medical test, part II, with plots



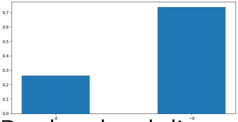
prior disease distribution



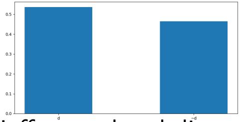
predicted test



observed test



Pearl-updated disease



Jeffrey-updated disease

Where we are, so far

Long introduction to probabilistic learning

Mathematical background

Expectation Maximisation (EM)

Conclusions

Distributions (finite, discrete)

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

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

- 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 $X \rightsquigarrow Y$, with special arrow.
- For $\sigma \in \mathcal{D}(X)$ and $c \colon X \rightsquigarrow Y$ 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:

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

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:

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

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

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

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

- The predicted test distribution is:

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

This gives the 13.5% likelihood of positive tests.

iCIS | Digital Security
Radboud University

Multisets (aka. bags)

- A multiset is a 'subset' in which elements may occur multiple times
  - for instance: $3|R\rangle + 2|G\rangle + 5|B\rangle$
  - in general: $\sum_i n_i|x_i\rangle$ of elements x_i with multiplicity $n_i \in \mathbb{N}$
- Typical examples:
  - coloured balls in an urn
  - votes per candidate in an election
  - solutions of a (polynomial) equation
  - data items, like age of study participants (in years)
- Frequentist learning turns a (non-empty) multiset into a distribution via normalisation:

$$Flrn\Big(\sum_i n_i|x_i\rangle\Big) := \sum_i \frac{n_i}{n}|x_i\rangle \qquad \text{where } n := \sum_i n_i.$$

- e.g. $Flrn\Big(3|R\rangle + 2|G\rangle + 5|B\rangle\Big) = \frac{3}{10}|R\rangle + \frac{2}{10}|G\rangle + \frac{5}{10}|B\rangle$.

Divergence between distributions/states

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

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

It is one standard way to compare distributions

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): $D_{KL}\big(c \ggg \omega,\ c \ggg \rho\big) \leq D_{KL}(\omega, \rho)$*

(4) *And: $D_{KL}\big(\omega \otimes \omega',\ \rho \otimes \rho'\big) = D_{KL}\big(\omega, \rho\big) + D_{KL}\big(\omega', \rho'\big)$*

iCIS | Digital Security
Radboud University

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 $c\colon X \rightsquigarrow Y$ and a predicate q on Y, one defines predicate transformation $c \mathrel{=\!\!\ll} q$, as predicate on X.

Explicitly, on $x \in X$,

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

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:

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

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:

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

It's the normalised product of ω and p.

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

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

Validity and conditioning example

- Take $X = \{1,2,3,4,5,6\}$ with state dice $\in \mathcal{D}(X)$
  - Explicitly: $dice = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$
- Take the predicate $evenish \colon X \to [0,1]$

$$\begin{array}{lll} evenish(1) = \frac{1}{5} & evenish(3) = \frac{1}{10} & evenish(5) = \frac{1}{10} \\ evenish(2) = \frac{9}{10} & evenish(4) = \frac{9}{10} & evenish(6) = \frac{4}{5} \end{array}$$

- The validity of $evenish$ for our fair dice is:

$$dice \models evenish = \sum_x dice(x) \cdot evenish(x) = \tfrac{1}{2}.$$

- If we take $evenish$ as evidence, we can update our $dice$ state and get:

$$\begin{aligned} dice\big|_{evenish} &= \sum_x \frac{dice(x)\cdot evenish(x)}{dice \models evenish} \, |x\rangle \\ &= \tfrac{1/6\cdot 1/5}{1/2}|1\rangle + \tfrac{1/6\cdot 9/10}{1/2}|2\rangle + \tfrac{1/6\cdot 1/10}{1/2}|3\rangle + \tfrac{1/6\cdot 9/10}{1/2}|4\rangle + \tfrac{1/6\cdot 1/10}{1/2}|5\rangle + \tfrac{1/6\cdot 4/5}{1/2}|6\rangle \\ &= \tfrac{1}{15}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{1}{30}|3\rangle + \tfrac{3}{10}|4\rangle + \tfrac{1}{30}|5\rangle + \tfrac{4}{15}|6\rangle. \end{aligned}$$

Two basic results about validity $\models$

Theorem (Validity and transformation)

For channel $c\colon X \rightsquigarrow Y$, 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 $c\colon X \rightsquigarrow Y$ 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"

$$Y \xrightarrow{c_\sigma^\dagger} X \qquad \text{with} \qquad c_\sigma^\dagger(y) \coloneqq \sigma|_{c \ll 1_y}$$

- 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
  - such a dagger / inversion is common in quantum theory

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

Set-up:

- a channel $c \colon X \to Y$ 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:
$$\sigma_P \coloneqq \sigma\big|_{c \ll q}$$

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

Back to the running disease-test example

Recall that we had 80% certainty of a positive test.

- **Pearl's update rule**
  (1) Evidence is predicate $q = \frac{4}{5} \cdot 1_p + \frac{1}{5} \cdot 1_n$,
  (2) Updated state:

$$\begin{aligned} \mathit{Pearl\text{-}posterior} \;\coloneqq\; \mathit{prior}|_{\mathit{test} \ll q} &= \tfrac{74}{281}|\, d \,\rangle + \tfrac{207}{281}|\, d^{\perp} \,\rangle \\ &\approx 0.26|\, d \,\rangle + 0.74|\, d^{\perp} \,\rangle \end{aligned}$$

- **Jeffrey's update rule**
  (1) Evidence is state $\tau = \frac{4}{5}|\, p \,\rangle + \frac{1}{5}|\, n \,\rangle$,
  (2) Updated state:

$$\begin{aligned} \mathit{Jeffrey\text{-}posterior} \;\coloneqq\; \mathit{test}^{\dagger}_{\mathit{prior}} \gg\!= \tau &= \tfrac{278}{519}|\, d \,\rangle + \tfrac{241}{519}|\, d^{\perp} \,\rangle \\ &\approx 0.54|\, d \,\rangle + 0.46|\, d^{\perp} \,\rangle \end{aligned}$$

Key results about Pearl & Jeffrey updates

Theorem

Let $c \colon X \rightsquigarrow Y$ 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,*

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

- Pearl is learning by encouragment, Jeffrey by discouragement
- The proof for Pearl is easy, but not for Jeffrey, see MFPS'21 paper

Where we are, so far

Long introduction to probabilistic learning

Mathematical background

Expectation Maximisation (EM)

Conclusions

EM background / set-up

- **inputs**:
  - a multiset ψ of data items on a set Y
  - a finite set X of classification labels
- **method:** determine
  - a mixture $\omega \in \mathcal{D}(X)$ of labels
  - a channel $c\colon X \to \mathcal{D}(Y)$, probabilistically mapping labels to data
- **goal:**
  - minimal divergence $D_{KL}\Big(Flrn(\psi),\, c \ggg\!= \omega\Big)$

In practice:

- the channel is of a parametrised class, written as $c[\theta]$
- the goal is hardly ever made explicit in the literature

EM, via iterations

- Recall, data multiset ψ is given, plus set X of labels.
- **Initialisation**: choose arbitrary $\omega^{(0)} \in \mathcal{D}(X)$ and parameter $\theta^{(0)}$; set $c^{(0)} := c[\theta^{(0)}] \colon X \rightsquigarrow Y$
- **E-step**: use Jeffrey's update rule in:

$$\omega^{(n+1)} := \left(c^{(n)}\right)^{\dagger}_{\omega^{(n)}} \gg\!= Flrn(\psi) \in \mathcal{D}(X)$$

- **M-step**: find minimal

$$\theta^{(n+1)} := \operatorname*{argmin}_{\theta} D_{KL}\Big(Flrn(\psi),\, c[\theta] \gg\!= \omega^{(n+1)}\Big)$$

(via solving a derivative-is-zero situation)

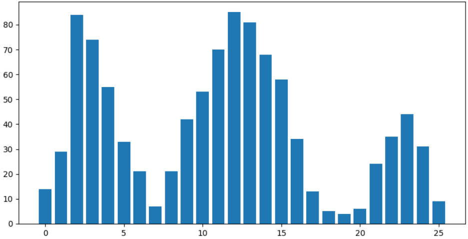
EM correctness

We get a decrease of divergence with each step:

$$
\begin{aligned}
& D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n+1)}] \gg\!= \omega^{(n+1)}\Big) \\
& \quad \leq D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n)}] \gg\!= \omega^{(n+1)}\Big) && \text{since } \theta^{(n+1)} \text{ is argmin} \\
& \quad \leq D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n)}] \gg\!= \big(c[\theta^{(n)}]^{\dagger}_{\omega^{(n)}} \gg\!= Flrn(\psi)\big)\Big) && \text{by defn of } \omega^{(n+1)} \\
& \quad \leq D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n)}] \gg\!= \omega^{(n)}\Big) && \text{by Jeffrey!}
\end{aligned}
$$

EM example

Consider the multiset of data over $\{0, 1, \ldots, 25\}$.



It consists of $N = 1000$ samples from the mixture of binomial distributions:

$$\tfrac{1}{2} \cdot bin[N]\big(\tfrac{1}{2}\big) \;+\; \tfrac{1}{3} \cdot bin[N]\big(\tfrac{1}{8}\big) \;+\; \tfrac{1}{6} \cdot bin[N]\big(\tfrac{9}{10}\big)$$

Aim: rediscover the mixture weights $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ and the biases $(\frac{1}{2}, \frac{1}{8}, \frac{9}{10})$.

EM example, continued

round	KL-div	mixtures $\omega^{(n)}$	biases $\theta^{(n)}$
0	0.853	$0.477\|1\rangle + 0.354\|2\rangle + 0.169\|3\rangle$	0.235, 0.389, 0.691
1	0.326	$0.353\|1\rangle + 0.35\|2\rangle + 0.297\|3\rangle$	0.159, 0.46, 0.754
2	0.132	$0.321\|1\rangle + 0.454\|2\rangle + 0.225\|3\rangle$	0.128, 0.478, 0.812
3	0.029	$0.311\|1\rangle + 0.515\|2\rangle + 0.174\|3\rangle$	0.122, 0.488, 0.872
4	0.011	$0.309\|1\rangle + 0.535\|2\rangle + 0.156\|3\rangle$	0.121, 0.493, 0.898

After 5 rounds we get pretty close to the original

- weights: $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$
- biases $\frac{1}{2}, \frac{1}{8}, \frac{9}{10}$

(The order is different, since labels are arbitrary)

Latent Dirichlet Allocation (LDA)

- LDA is a probabilistic algorithm for topic modeling
  - **input**:
    - several documents, as multisets of words
    - a set of topics
  - **output**: channels
    - $Doc \to \mathcal{D}\big(Top\big)$
    - $Top \to \mathcal{D}\big(Wrd\big)$
- The algorithm also works iteratively
  - the crucial role of Jeffrey's rule is identified in the paper

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 "XAI" (explainable AI)
- The two update rules of Pearl and Jeffrey:
  - can give wildly different outcomes — but agree on point evidence
  - are not so clearly distinguished in the literature — probably because fuzzy / soft predicates are not standard
  - Pearl increases validity, Jeffrey decreases divergence
  - the answers are "exclusive", see paper: Pearl need not decrease divergence, and Jeffrey need not increase validity
- Jeffrey's role is made explicit in basic machine learning algorithms EM and LDA
- Overal picture about Pearl versus Jeffrey remains unclear
  - impression: in statistics, Jeffrey is used, unless there is a conjugate prior situation. The fascination remains.

Thanks for your attention!

For much more info, see my book-in-the-making:

Structured Probabilistic Reasoning

http://www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf