The relation between leave-one-out log-scores and log-evidence

P.G.L. Porta Mana
<piero.mana@ntnu.no>

Draft of 8 August 2019 (first drafted 8 August 2019)

***

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1 ***

The probability calculus tells us unequivocally how our degree of belief in a hypothesis H_h given data D and knowledge K is related to our degree of belief in observing those data when we entertain that hypothesis as true:

$$\mathrm{P}(H_h \mid D\, K) = \frac{\mathrm{P}(D \mid H_h\, K)\, \mathrm{P}(H_h \mid K)}{\sum_h \mathrm{P}(D \mid H_h\, K)\, \mathrm{P}(H_h \mid K)}. \tag{1}$$

This post-data degree of belief is of course conditional on the set $\{H_h\}$ of mutually exclusive and exhaustive hypotheses under consideration (implicit in the knowledge K).

We can combine the post-data beliefs in two hypotheses in different ways to form a quantitative ideas of how much the data is giving evidence for one or the other. For example their ratio or the logarithm of their ratio, often called *weight of evidence* and *Bayes factor* (Good 1950; Osteyee et al. 1974; MacKay 1992; Kass et al. 1995; see also Jeffreys 1983 chs V, VI, A). 'It is historically interesting that the expression "weight of evidence", in its technical sense, anticipated the term "likelihood" by over forty years' (Osteyee et al. 1974 § 1.4.2 p. 12).

The literature in probability and statistics has also employed for quite some time various other ad-hoc measures to assess our beliefs in a set of hypotheses given some data. Here I consider one in particular, the *leave-one-out cross-validation log-score* (Krnjajić et al. 2011; Geisser et al. 1979; Vehtari et al. 2002; 2012; Draper et al. 2014; Piironen et al. 2017):

$$\frac{1}{d}\sum_{i=1}^{d} \ln \mathrm{P}(D_i \mid D_{-i}\, H_h\, K) \tag{2}$$

where every D_i is one datum in the data $D \equiv \bigwedge_i D_i$, and D_{-i} denotes the data with datum D_i excluded. The intuition behind this measure is more or less this: 'let's see what my belief in one datum should be, on average, once I've observed the other data, if I consider H_h as true'. 'On average' means considering such belief for every single datum in turn, and then taking the geometric mean (arithmetic mean on a log scale).

This is a reasonable intuition, and the log-score (2) and post-data probability (1) often lead to qualitatively similar results in comparing two hypotheses. There are exceptions, though.

My point of view, which hinges on the logical foundations of the probability calculus (Pólya 1941; 1949; 1968; Cox 1946; Hailperin 1996; Jaynes 2003; Paris 2006; Snow 1998; Terenin et al. 2017), is that every intuitively built quantitative assessment of belief is either (1) an approximation of a formula that can be derived from the probability calculus, or (2) wrong.

I shall now show that the log-score above can be viewed as an approximation of the logarithm of the post-data probability (1); or, if you like, that the post-data probability can be seen as a refined version of the log-score.

We can obviously write

$$\mathrm{P}(D \mid H\, K) \equiv \bigl[\underbrace{\mathrm{P}(D \mid H\, K) \times \dots \times \mathrm{P}(D \mid H\, K)}_{d\ \text{times}}\bigr]^{1/d} \tag{3}$$

where we have dropped the subscript $_h$ for simplicity. By the rules of probability we have

$$\mathrm{P}(D \mid H\, K) = \mathrm{P}(D_i \mid D_{-i}\, H_h\, K) \times \mathrm{P}(D_{-i} \mid H_h\, K) \tag{4}$$

and this holds no matter what specific $i \in \{1, \dots, d\}$ we choose (temporal ordering and similar matters are completely irrelevant in the formula above: it's a logical relation between propositions). So let's expand each of the d factors in the identity (3) using the product rule (4), but using a different datum in each of them. The result can be displayed thus:

$$
\begin{aligned}
\mathrm{P}(D \mid H\,K) \equiv \bigl[\, & \mathrm{P}(D_1 \mid D_{-1}\,H\,K) \times \mathrm{P}(D_{-1} \mid H\,K) \times \\
& \mathrm{P}(D_2 \mid D_{-2}\,H\,K) \times \mathrm{P}(D_{-2} \mid H\,K) \times \\
& \qquad\qquad \dots \qquad\qquad\qquad \times \\
& \underset{\substack{\uparrow \\ \text{log-score}}}{\mathrm{P}(D_d \mid D_{-d}\,H\,K)} \times \mathrm{P}(D_{-d} \mid H\,K) \,\bigr]^{1/d} .
\end{aligned}
\tag{5}
$$

Note that upon taking the logarithm of this expression, the factors vertically aligned on the left turn together into the log-score (2), as indicated.

Now note that steps we just made for $\mathrm{P}(D \mid H\,K)$ – the root-product identity (3) and the expansion (5) – can be done for each of the remaining factors $\mathrm{P}(D_{-i} \mid H\,K)$, and so on recursively. Here is an explicit example for $d = 3$:

$$
\begin{aligned}
&\mathrm{P}(D \mid H\,K) \equiv \\
&\quad \Bigl\{ \mathrm{P}(D_1 \mid D_2\,D_3\,H\,K) \times \bigl[\mathrm{P}(D_2 \mid D_3\,H\,K) \times \mathrm{P}(D_3 \mid H\,K) \times \\
&\qquad\qquad\qquad\qquad \mathrm{P}(D_3 \mid D_2\,H\,K) \times \mathrm{P}(D_2 \mid H\,K)\bigr]^{1/2} \times \\
&\quad\; \mathrm{P}(D_2 \mid D_1\,D_3\,H\,K) \times \bigl[\mathrm{P}(D_1 \mid D_3\,H\,K) \times \mathrm{P}(D_3 \mid H\,K) \times \\
&\qquad\qquad\qquad\qquad \mathrm{P}(D_3 \mid D_1\,H\,K) \times \mathrm{P}(D_1 \mid H\,K)\bigr]^{1/2} \times \\
&\quad\; \mathrm{P}(D_3 \mid D_1\,D_2\,H\,K) \times \bigl[\mathrm{P}(D_1 \mid D_2\,H\,K) \times \mathrm{P}(D_2 \mid H\,K) \times \\
&\qquad\qquad\qquad\qquad \mathrm{P}(D_2 \mid D_1\,H\,K) \times \mathrm{P}(D_1 \mid H\,K)\bigr]^{1/2} \Bigr\}^{1/3}
\end{aligned}
\tag{6}
$$

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Cox, R. T. (1946): *Probability, frequency, and reasonable expectation*. Am. J. Phys. **14**^{1}, 1–13. http://jimbeck.caltech.edu/summerlectures/references/ProbabilityFrequencyReasonableExpectation.pdf, https://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Cox_1946.pdf.

Draper, D., Krnjajić, M. (2014): *Bayesian model comparison: log scores and* DIC. Stat. Probab. Lett. **88**, 9–14.

Geisser, S., Eddy, W. F. (1979): *A predictive approach to model selection*. J. Am. Stat. Assoc. **74**^{365}, 153–160.

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Kass, R. E., Raftery, A. E. (1995): *Bayes factors*. J. Am. Stat. Assoc. **90**^{430}, 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf.

Krnjajić, M., Draper, D. (2011): *Bayesian model specification: some problems related to model choice and calibration*. http://hdl.handle.net/10379/3804.

MacKay, D. J. C. (1992): *Bayesian interpolation*. Neural Comp. **4**^{3}, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

Osteyee, D. B., Good, I. J. (1974): *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*. (Springer, Berlin).

Paris, J. B. (2006): *The Uncertain Reasoner's Companion: A Mathematical Perspective*, reprint. (Cambridge University Press, Cambridge). See also Snow (1998).

Piironen, J., Vehtari, A. (2017): *Comparison of Bayesian predictive methods for model selection*. Stat. Comput. **27**^{3}, 711–735.

Pólya, G. (1941): *Heuristic reasoning and the theory of probability*. Am. Math. Monthly **48**^{7}, 450–465. First written in French 1939.

— (1949): *Preliminary remarks on a logic of plausible inference*. Dialectica **3**$^{1–2}$, 28–35.

— (1968): *Mathematics and Plausible Reasoning: Vol. II: Patterns of Plausible Inference*, 2nd ed. (Princeton University Press, Princeton). First publ. 1954.

Snow, P. (1998): *On the correctness and reasonableness of Cox's theorem for finite domains*. Comput. Intell. **14**^{3}, 452–459.

Terenin, A., Draper, D. (2017): *Cox's theorem and the Jaynesian interpretation of probability*. arXiv:1507.06597. First publ. 2015.

Vehtari, A., Lampinen, J. (2002): *Bayesian model assessment and comparison using cross-validation predictive densities*. Neural Comp. **14**^{10}, 2439–2468.

Vehtari, A., Ojanen, J. (2012): *A survey of Bayesian predictive methods for model assessment, selection and comparison*. Statist. Surv. **6**, 142–228.