Prediction error weighting

Johan Kwisthout

October 24, 2017

1 Introduction

It is often assumed in the predictive processing account that prediction errors are weighted: the more a prediction error can be used to update generative models, the higher this weight should be. In the literature this is often referred to as *precision-weighted* prediction errors; the prediction error is weighted according to the expected precision of the prediction, that is, how much noise there will be in the signal. The noisier the signal is expected to be, the less influence a prediction error should have.

When we use causal Bayesian networks as a computational framework for realizing predictive processing, this concept is not as usable. We propose that generative models are based on hyperpriors, that capture the *confidence* in a particular probability distribution. Specifically, for discrete distributions the hyperpriors are Dirichlet distributions $f(x; \alpha_1, \ldots, \alpha_n)$ with hyperparameters $\alpha_1, \ldots, \alpha_n$, where n is the number of parameters in the (original, discrete) distribution plus one. In this paper we use (without loss of generality) $n = 2$ (for binary distributions, that have just a single parameter) and refer to them as α and β. We assume that there is a generative model $\Pr(X)$, defined by the current values of α and β such that $\Pr(X = x) = \frac{\alpha}{\alpha+\beta}$ and $\Pr(X = \bar{x}) = \frac{\beta}{\alpha+\beta}$. We assume that the current observation is $X = x$ and we define the prediction error $\mathrm{D_{KL}}(\mathrm{Pr_{Pred}} \| \mathrm{Pr_{Obs}})$ as the KL-divergence between the distribution $\mathrm{Pr_{Pred}}$ as defined above and the distribution $\mathrm{Pr_{Obs}}$ as $\Pr(X = x) = 1$. This KL-divergence is defined as

$$\mathrm{D_{KL}}(\mathrm{Pr_{Obs}} \| \mathrm{Pr_{Pred}}) = \sum_{\mathbf{p} \in \Omega(\mathrm{Obs})} \mathrm{Pr_{Obs}}(\mathbf{p}) \ln \left(\frac{\mathrm{Pr_{Obs}}(\mathbf{p})}{\mathrm{Pr_{Pred}}(\mathbf{p})} \right)$$

where the term $0 \ln 0$ as 0 when appearing in this formula is interpreted as 0 as $\lim_{x \to 0} x \ln x = 0$. Note that we here define the KL-divergence in nats, rather than in bits, using the natural logarithm rather than $\log_2$.

We (normatively) define the precision-weighted prediction error as the *effect of the observation on the underlying hyperprior*. That is, on observation of an event associated with α, we compute the (continuous) KL-divergence between $f(x; \alpha, \beta)$ and $f(x; \alpha+1, \beta)$. Alternatively, one can refer to 'weighted prediction error' as $W \times \mathrm{D_{KL}}(\mathrm{Pr_{obs}} \| \mathrm{Pr_{pred}})$. Can W be analytically derived in terms of α

and β? The answer to this question turns out to be 'yes'. Moreover, we show that our normative definition of precision-weighted prediction errors captures exactly these properties of weights that we would like to have.

2 Derivation

Let $\mathrm{Pr}_{\mathrm{Pred}}(X = x) = \frac{\alpha}{\alpha+\beta}$, $\mathrm{Pr}_{\mathrm{Pred}}(X = \bar{x}) = \frac{\beta}{\alpha+\beta}$, $\mathrm{Pr}_{\mathrm{Obs}}(X = x) = 1$, and $\mathrm{Pr}_{\mathrm{Obs}}(X = \bar{x}) = 0$, for unspecified $\alpha \geq 1$, $\beta \geq 1$. We then have that $\mathrm{D}_{\mathrm{KL}}(\mathrm{Pr}_{\mathrm{Obs}} \| \mathrm{Pr}_{\mathrm{Pred}}) = 1 \times \ln({}^{1}/_{\frac{\alpha}{\alpha+\beta}}) + 0 \times \ln({}^{0}/_{\frac{\beta}{\alpha+\beta}}) = \ln(\frac{\alpha+\beta}{\alpha})$. For the KL-divergence between $f'(x; \alpha+1, \beta)$ and $f(x; \alpha, \beta)$ we use the approach demonstrated by Bariç Kurt[1]. We use that $\mathrm{D}_{\mathrm{KL}}(\mathrm{f'(x)} \| \mathrm{f(x)}) = \left\langle \ln(\frac{f'(x)}{f(x)}) \right\rangle_{f'(x)} = \langle \ln f'(x) - \ln f(x) \rangle_{f'(x)}$ and expand and re-arrange this geometric mean:

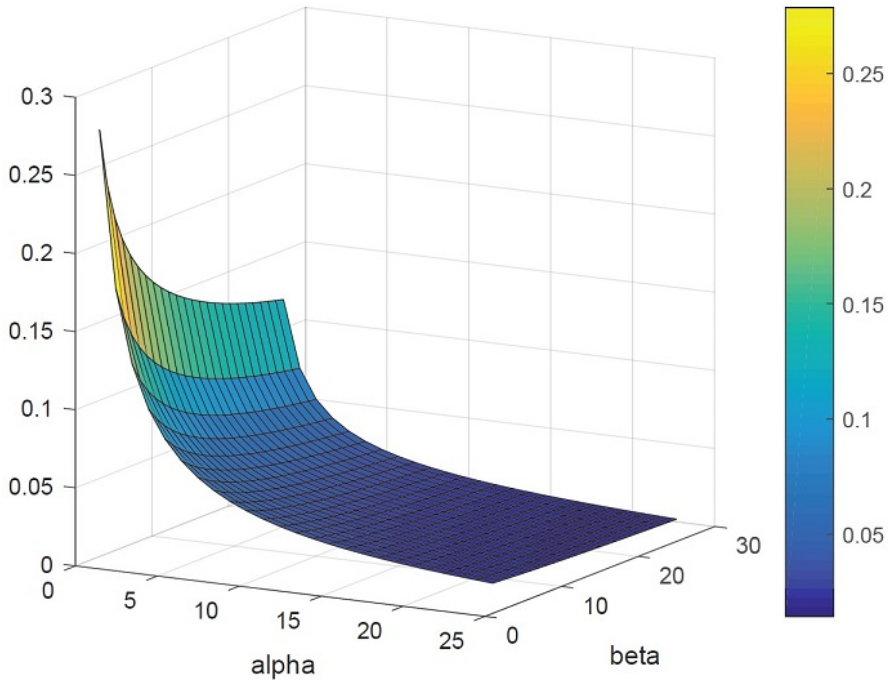
$$
\begin{aligned}
&\langle \ln f'(x) - \ln f(x) \rangle_{f'(x)} && = \\
&\langle \ln\Gamma(\alpha+\beta+1) - \ln\Gamma(\alpha+1) - \ln\Gamma(\beta) + (\alpha+\beta)\ln x \\
&\qquad - \ln\Gamma(\alpha+\beta) + \ln\Gamma(\alpha) + \ln\Gamma(\beta) - (\alpha+\beta-1)\ln x \rangle_{f'(x)} && = \quad (1) \\
&\ln\Gamma(\alpha+\beta+1) - \ln\Gamma(\alpha+\beta) + \ln\Gamma(\alpha) - \ln\Gamma(\alpha+1) + \langle \ln x \rangle_{f'(x)} && = \quad (2) \\
&(\ln(\alpha+\beta)! - \ln(\alpha+\beta-1)!) - (\ln(\alpha)! - \ln(\alpha-1)!) \\
&\qquad + \psi(\alpha+1) - \psi(\alpha+\beta+1) && = \quad (3) \\
&\ln(\alpha+\beta) - \ln\alpha + \psi(\alpha+1) - \psi(\alpha+\beta+1) && = \quad (4) \\
&\ln(\alpha+\beta) - \ln\alpha + (H_\alpha - \gamma) - (H_{\alpha+\beta} - \gamma) && = \quad (5) \\
&\ln(\alpha+\beta) - \ln\alpha + H_\alpha - H_{\alpha+\beta} && \quad (6)
\end{aligned}
$$

We used in Step 3 that $\langle \ln x \rangle_{f'(x)} = \psi(\alpha+1) - \psi(\alpha+\beta+1)$ and $\Gamma(x) = (x-1)!$. In Step 4, H_n is the n-th harmonic number, where γ is the Euler-Mascheroni constant. Note that, since $\lim_{x\to\infty} H_x = \ln x$, in the limit the weighted prediction error approaches zero. We thus have $W \times \mathrm{D}_{\mathrm{KL}}(\mathrm{Pr}_{\mathrm{Obs}} \| \mathrm{Pr}_{\mathrm{Pred}}) = \ln(\frac{\alpha+\beta}{\alpha}) + H_\alpha - H_{\alpha+\beta}$ and thus $W = 1 + \frac{H_\alpha - H_{\alpha+\beta}}{\ln(\frac{\alpha+\beta}{\alpha})}$. W has its maximum for $\alpha = \beta = 1$ and decreases to zero with increasing α and β as shown on the following figure.

3 Conclusion and Future work

We defined *precision-weighted prediction error* as the KL-divergence between the hyperprior, representing a parameter in a generative model, before and after updating the hyper-parameters as a consequence of observing an event. We showed that in this definition the 'weight' of the prediction error (defined as the KL-divergence between prediction and observation) can be analytically derived and has the desired properties. This model does not yet address the following theoretical questions:

[1] `http://bariskurt.com/kullback-leibler-divergence-between-two-dirichlet-and-beta-distributions`.



- What if we did not observe an event with full certainty; how should then the hyperparameters be updated?
- The updating of generative models here does not take into account *model revision* or *model refinement*; this is pure model updating.
- We simplified a 'generative model' and collapsed it to be a single variable. When models consist of non-trivial interactions between multiple variables, such as multiple causes for the same effect, the question is: *What* part of the prediction error accounts for *which* parameter update.