A delta rule approximation to Bayesian inference in change-point problems

Robert C. Wilson, Matthew R. Nassar and Joshua I. Gold

INTRODUCTION

How does the brain make useful inferences in a rapidly changing world? For example, what will be the next value in this change-point problem?

OPTIMAL INFERENCE

If change-point locations are given

Inference is based only on data points up to last change-point. $x_t$ $r_t$

The number of samples since the last change-point is called the run-length, $r_t$, with very simple dynamics

\[ r_{t+1} = \begin{cases} r_t + 1 & \text{no change-point} \\ 0 & \text{change-point} \end{cases} \]

If change-point locations are unknown [1, 2, 3]

Maintain distribution over run-lengths given data $p(r_t|x_{1:t})$ $r_t$

If the rate at which change-points occur, $h$, is given, this is computed recursively with message passing on a graph

\[ p(r_t|x_{1:t}) \propto \sum p(r_t|r_{t-1})p(x_t|r_{t-1})p(r_{t-1}|x_{1:t-1}) \]

Predictive distribution is then computed as marginal over run-length:

\[ p(x_{t+1}|x_{1:t}) = \sum_{r_t} p(x_{t+1}|r_t)p(r_t|x_{1:t}) \]

Example output of optimal model

PROBLEM WITH OPTIMAL INFERENCE

Possible values for run-length grow linearly with time and are unbounded. It seems unlikely that the brain can represent this distribution.

REDUCED MODEL

Build an approximation based on just two possible values for the run-length $r_t = 0$ and $r_t = \hat{r}_t$

\[ p(r_{t-1}|x_{1:t-1}) = (1-h)\delta(r_{t-1} = \hat{r}_{t-1}) + h\delta(r_{t-1} = 0) \]

Update rule now has two stages

A: Expansion. Update as before to get a distribution over three possible values of run-length

\[ p^A(r_t|x_{1:t}) = (1-h)(1-p_t^{ch})\delta(r_t = \hat{r}_{t-1} + 1) + (1-h)p_t^{ch}\delta(r_t = 1) + h\delta(r_t = 0) \]

where $p_t^{ch}$ is the probability of change on the last trial

\[ p_t^{ch} = \frac{hp(x_t|r_{t-1} = \hat{r}_{t-1})}{(1-h)p(x_t|r_{t-1} = \hat{r}_{t-1}) + (1-h)p(x_t|r_{t-1} = 0) \}

B: Contraction. Reduce the three-valued $p^A(r_t|x_{1:t})$ to a two-valued distribution

\[ p^B(r_t|x_{1:t}) = (1-h)\delta(r_t = \hat{r}_t) + h\delta(r_t = 0) \]

Such that in some sense

\[ p^B(r_t|x_{1:t}) \approx p^A(r_t|x_{1:t}) \]

This update rule also has a graphical interpretation:

\[ \text{Contraction is achieved by matching moments} \]

Say $p^B(r_t|x_{1:t}) \approx p^A(r_t|x_{1:t})$ when the first $M$ moments of the two are matched; i.e., for $m = 1, 2, \ldots, M$

\[ \langle x_{1:t}^m \rangle p(x_{t+1}|r_t = \hat{r}_t) = (1-p_t^{ch}) \langle x_{1:t}^m \rangle p(x_{t+1}|r_t = \hat{r}_{t-1} + 1) + p_t^{ch} \langle x_{t+1}^m \rangle p(x_{t+1}|r_t = 0) \]

For exponential-family distributions this turns out to be all we need to recover $\hat{r}_t$

DELTA RULE FOR CHANGE-POINTS

With moment matching, the mean, $\mu_t$, of the predictive distribution updates according to the following delta rule

\[ \mu_{t+1} = \mu_t + \alpha_t (x_t - \mu_t) + \beta_t (\mu_0 - \mu_t) \]

with

\[ \alpha_t = \frac{1 - p_t^{ch}}{\hat{r}_{t-1} + v_0 + 1} + p_t^{ch} \frac{v_0}{v_0 + 1} \quad \text{and} \quad \beta_t = \frac{v_0 p_t^{ch}}{v_0 + 1} \]

$\alpha_t$ is the learning rate and determines the extent to which new information influences current beliefs, and $\beta_t$ determines the rate at which the predictive mean regresses to the prior mean, $\mu_0$. $v_0$ is a constant, the “equivalent sample size” of the prior.

This delta-rule is very efficient to implement and biologically much more plausible [e.g., 4] than the full, optimal model.

EXAMPLE: GAUSSIAN WITH CHANGING MEAN Optimal model:

Reduced model:

REFERENCES